LF AI & DATA

LF AI & DATA GENERATIVE AI COMMONS

責任ある生成AI フレームワーク（RGAF） — V0.9

※ RGAF：Responsible Generative AI Framework

The Linux Foundation AI and Data Foundation, Generative AI Commons

2025年3月

著者は、Haluk Demirkan、Adel Zaalouk、Suparna Bhattacharya、Susan Malaikaであり、LF AI and Data の Generative AI Commonsにおける Responsible AI Workstreamのメンバー Karen Bennet、Imani Carey、Oita Coleman、David Edelsohn、David Ellison、Andreas Fehlner、Martin Foltin、Adrian Gonzales、Ali Hashmi、Ofer Hermoni、Vini Jaiswal、Anni Lai、Kevin Lu、Victor Lu、Raghavan Muthuregunathan、Michael Novak、Ronald Petty、Jeff Polack、Annmary Roy、Santhosh Sachindranより、貴重なフィードバック、サポート、貢献を受けました。

エグゼクティブ サマリー

認知の時代（The Cognitive Age）、すなわち、精神と機械が協力し合う第5次産業革命（5IR）[1]は、社会の構造や人間の存在そのものを再定義する態勢が整った人類の歴史におけるデジタル変革の転換点にあります。人工知能（AI）を活用したスマート テクノロジー ソリューション、ハードウェア デバイス、そしてブレイン コンピューター インターフェースの能力は、絶えず進化を続け、私たちの生活に不可欠な要素となりつつあります。AIの機能は、移動性、機敏さ、運動制御から、多言語音声認識、自然言語翻訳、新しいタスクの学習、そして人間の五感（触覚、視覚、聴覚、嗅覚、味覚）を通じた物理世界とのやり取りまで、急速に向上しています。場合によっては、パターン認識、構造化問題と非構造化問題の両方の解決、文書作成、絵画、映画制作、作曲などにおいて、人間よりもはるかに高速かつ優れた能力を発揮します[2]。

これらのソリューションは、コラボレーション、グローバル化、自動化、そして特定のタスクの迅速な完了といった事例を生み出す一方で、悪意のある人物による虚偽情報の作成（例えば、フェイク ニュース、非人間的な意思決定、プライバシーや倫理の侵害）に利用されるリスクも伴います。共感性、人間的な感覚や感触の欠如、そしてタスク以外の環境への注意力の欠如により、破壊行為を行うようにプログラムまたは訓練される可能性があります。人間と技術の価値共創を考慮しなければ、これらのスマート ソリューションは人間にとって最適ではない結果をもたらす可能性があります。

責任あるAIは、有益なAIの開発と展開における倫理、道徳、信頼、法的価値を網羅し、長期的な効率性と有効性を実現するAIガバナンスの新たな分野です。AIソリューションが多様なデータセットを用いて設計･学習され、公平性を確保するためにモデル化されるには、幅広いステークホルダーがオープンソースの考え方に基づいて、ガバナンスに関する議論に参加する必要があります。私たちは既に、創造性、イノベーション、競争、より安全なAI、そして透明性の向上を育むために、オープンソースの考え方と開発を促進することの価値を実感しています[3]。Linux Foundation AI and Data Generative AI Commonsによる責任ある生成AIフレームワーク（RGAF：Responsible Generative AI Framework）は、9つの基本的視点のバランスを取りながら、個人と社会の利益のためにAIソリューションを倫理的かつ意図的に設計、開発、展開できるように組織を導くことを目的としています。主なポイントとして、以下のベスト プラクティスが浮かび上がりました。

1. RGAFを活用することで、オープンソースの生成AIプロジェクトにおける複雑な責任あるAIの課題を乗り越えることができます。RGAFの9つの視点は、EU AI法、NIST AIフレームワーク、シンガポールAI戦略、中国のAI開発など、主要なグローバルAIフレームワークの要素と整合しています。
2. RGAFの9つの視点、それぞれの視点に対応する課題、そして実践的な改善策を理解できます。9つの視点とは、人中心と人との整合性、アクセス性と包摂性、堅牢性、信頼性と安全性、透明性と説明可能性、説明責任と是正可能性、プライバシーとセキュリティ、遵守性と制御可能性、倫理性と公正性、環境的持続可能性です。
3. Model Openness Framework（MOF）を採用し、透明性と再現性を促進し（例えば、オープン データセット、コード、ドキュメントを通じて）、RGAF原則の独立した検証を促進します。
4. Decoding Trustなどの評価フレームワークを採用し、AIモデルの信頼性を定量化し、RGAF原則との整合性を確保します。
5. AIモデルだけでなく、複合AIシステムのあらゆるコンポーネントにおいて、RGAFのすべての視点が評価され、実践において一貫して適用されるようにします。**キーワードは、**人工知能、生成AI、知能拡張、人間参加型AI（HITL：Human In The Loop）、スマート コンピューター、オープンソース、アウトタスキング、プロセス自動化です。

1. はじめに

第五次産業革命は、飛躍的に成長する技術力、バリューチェーン、バリューショップ、バリューネットワークを基礎とした第四次産業革命の基盤の上に成り立つものです。デジタル技術による革命のスピードは前例のないものであり、その影響はより多くの人々に、より多様な形で、関わってきます。この新たな時代である第五次産業革命の瀬戸際に立つ今、人類の進歩にとって、人間と機械の知能の調和が不可欠となっています。それは、言語学、脳理論、人間行動学、神経学、哲学、社会学、心理学、数学、認知科学、サービス科学、コンピューター科学、経営学、デザイン科学、意思決定科学、データ科学、そして機械学習、ロボット工学、IoTといった技術がプロセスに統合され、複雑で適応性の高いサービス システムを構築することを特徴としています[4]。

組織は現在、業務効率の向上と価値創造を実現する新たなソリューションの開発を目指し、AIと生成AIに多額の投資を行っています。顧客サービス エージェントとしてのチャットボット、ソフトウェア開発エンジニアとしてのアプリケーション、そしてテキスト、画像、動画、音声、音楽生成のためのソリューションを展開しています。テキスト生成のためのChat GPTやBERT、画像生成のためのDALL-EやStyleGan、音声・音楽生成のためのJukeboxやWaveNet、Microsoftアプリケーションのための Copilot、検索とナレッジマネジメントのためのGoogle AI、コード生成、複数のデータソース、リポジトリ、企業情報システムの統合のためのAmazon Q、音声コマンドによるAlexa、Siri、Cortanaなど、私たちは日常生活の中で、仮想アシスタントによる迅速な回答や支援を求め、生成AIを利用しています。生成AIは、組織、業界、そして社会を変革する破壊的な力となりました。McKinseyによると[5]、Gen AIはソフトウェア市場の大部分に破壊的な変化をもたらし、顧客業務、マーケティング、営業、研究開発、ソフトウェア エンジニアリングに大きな影響を与える可能性があります。GenAIはソフトウェア分野の大幅な成長を牽引するでしょう。2027年までに、この技術への支出は1,750億ドルから2,500億ドルに達し、この分野の成長を2%～6%押し上げることに貢献しています。生成AIは、様々な分野で広く成功を収めており、大きな期待が寄せられています。しかし同時に、これらのスマート ソリューションが情報の偽装、プライバシー侵害、著作権侵害、法的リスクの侵害、有害コンテンツの配信、機密情報の漏洩、そして悪意のある人物の新たなサイバー攻撃への備えといった可能性に対する懸念も高まっています。

最も一般的に使用されているGenAIモデルはクローズドソース、または「プロプライエタリ」とも呼ばれ、所有者がライセンスを許可しているため、事実上、一般公開されている私有財産です。また、時には「ブラックボックス」と見なされることもあります。つまり、その仕組みを知っているのは開発者だけなので、内部で何が起こっているのかを正確に把握することが困難であることを意味しています。これは通常、商業的な理由によるものです。開発者は販売することで収益を得ており、もし誰もが仕組みを知ってしまえば、再現して販売（または無償提供）できるからです。クローズドソースAIツールはアクセスしやすく、使いやすく、信頼性の高い保守とサポートが期待されるため、エンドユーザーにとって多くのメリットがあります。ユーザーはまた、これらのソリューションがデータ漏洩や不正アクセスに対して高度なセキュリティを備えていることを期待しています。もしソリューションにこのような備えを期待しているのであれば、そこには、深刻な評判の失墜や、データ保護法に基づく罰金のリスクがあります。このようなユーザーは、アップデートをベンダーに依存し、カスタマイズ オプションが限られていることを意味するからです。特に、ベンダーがカスタム バージョンを提供するビジネス ケースが少ないニッチ市場では、その傾向が顕著です。GPT-4、GoogleのGemini、画像モデルのDall-EとMidjourney、Nvidia Jarvisはすべて、クローズドソースの生成AIモデル例です。

GenAIの透明性と再現性の欠如は、導入の進行を妨げ、信頼を損ないます。Linux Foundationが2023年12月に発表した「2023年オープンソース生成AI調査レポート」[6]では、次のように報告されています。

「透明性、コラボレーション、そしてイノベーションの共有という原則に根ざしたオープンソースのアプローチは、GenAI技術の発展に変革をもたらす可能性を秘めています。AIアルゴリズムとデータセットへのアクセスを民主化することで、オープンソースの取り組みは、幅広く多様な開発者がGenAIシステムに貢献し、改良し、批評することを可能にします。こうした集合知はイノベーションの速度を加速させ、閉鎖的な開発環境では見過ごされがちなバイアスや脆弱性を発見し、修正します。」

クローズドソース モデルとは異なり、オープンソース モデルでは、開発者は「覗き見」し、その仕組みを理解することができます。これによりモデルを改善したり、新しいタスクやユース ケースに適応させたりする機会を見つけることができるかもしれません[7]。セキュリティの観点から見ると、オープンソース モデルは定義上、外部監査を受けることができ、セキュリティ上の欠陥が開発者コミュニティによって発見され、（願わくは）修正されることが保証されます。ビジネスの観点から見ると、オープンソース モデルの最大の利点は、少なくとも理論上は、新しいアプリケーションやサービスの開発に実質的に無料で使用できることにあると言えるでしょう。実際には、セットアップや期待どおりに動作させるには、多くの場合、費用がかかります。こうしたサポートはコミュニティから無料で提供される場合もありますが、サードパーティの商用プロバイダーとの契約が必要になる場合もあります。オープンソースの生成AIモデルの最も有名な例の一つは、最も人気のあるテキスト画像生成ツールの一つであるStable Diffusionです。もう一つの例はMetaのLlamaです。これは、ChatGPTを支えるOpenAIのクローズドソースGPTモデルの代替として機能する言語モデルです。

データの偏り、ディープ フェイクの安全性、AI生成コンテンツに関する法的問題への懸念が高まる中、スマート テクノロジーを活用したソリューションを設計、開発、展開するためのフレームワークが重要です。GenAIを産業用途に導入するにあたり、新たな取り組みと考慮事項が必要です[8]。これらのフレームワークは、合成モデル生成やディープ フェイクといったGenAIモデルの出力に関連する潜在的な安全性の懸念に対処できます。また、重要な意思決定が依然として人間によって行われていることを保証するために、人間による監視に関するガイドラインを定義するのにも役立ちます。さらに、知的財産権やAI生成コンテンツの説明責任といった問題に対処するために、法的および規制上の枠組みの進化も必要です。Googleなどの企業は、検出技術の活用、専門チームの育成、有害コンテンツの拡散阻止に向けた業界をリードする取り組みなどを通じて、信頼と安全のためのソリューションを構築しています[9]。TikTokは、AI生成コンテンツ（AIGC：AI Generated Content）に関するクリエイター向けの公開ガイドライン[10]を策定し、クリエイターがAI生成コンテンツ（AIGC：AI-generated content）を使用して安全かつ責任を持って創造性を表現できるようにするための継続的な取り組みの最新情報を共有しています[11]。

AIに関連する害悪について人々は懸念を表明しており[12]、その懸念は技術のライフサイクル、つまりその構築方法から現実世界での応用方法まで多岐にわたります。以下に主な懸念事項を示します。

- **AIがどのように構築されたか。**著作権と知的財産権です。すなわち、トレーニングおよび活用において、同意なしにコンテンツを使用されることです。生体認証データや音声、顔、指紋、等の個人データは悪用や詐欺につながる可能性があります[13]。これには合成データも含まれます。
- **現実感の欠如。**合成データは現実世界のデータの複雑さやニュアンスを欠く可能性があり、AIモデルが現実世界のシナリオで十分なパフォーマンスを発揮できない可能性があります。合成データのみで学習したAIモデルは、合成データと実際のデータとの差異により、現実世界の状況に効果的に一般化できない可能性があります[14] [15]。
- **AIがどのように機能し、ユーザーと対話するか。**バイアスや不正確さです。すなわち、ユーザーのアイデンティティ分析が、ユーザーの経済状況、生計や雇用機会に影響を与える意思決定につながる可能性があることです[16]。
- **透明性と説明可能性の欠如。**AIシステムは非常に複雑で理解しにくいため、ユーザーが意思決定の方法を理解し、評価することが困難です。
- **悪意ある行為者。**ディープ フェイクは、敵対的生成ネットワーク（GAN：Generative Adversarial Network）などのディープラーニング技術を用いて、実在の人物をデジタル的に改変し、模倣します。悪質な例としては、上司から従業員への指示を模倣したり、苦境にある家族に偽のメッセージを送ったり、個人の恥ずかしい写真を配布したりすることが挙げられます。国土安全保障省（Homeland Security）の調査では、ディープ フェイクのような障害を克服するために、官民の関係者間の緊密な協力が求められています[17]。
- **人は関与ませんが、しかし。**人間のAIシステムへの依存は、システムが故障したり、侵害されたり、廃止されたりした場合に、悪影響を引き起こす可能性があります[18]。
- **倫理。**例えば、大規模監視に向けた顔認証技術の利用、健康管理や自動運転の設定における生死に関わる意思決定を行うAIの利用です。これら人の音声、顔、指紋、個人データなどは、悪用や詐欺につながる可能性があります[19]。
- **詐欺、不正行為、悪意ある利用。**詐欺師が生成AI製品を使ってフィッシング メールを作成するようになり、以前はあった同音意義単語の様なスペルミスや文法ミスが無くなったため、フィッシングメールを見分けるのが難しくなるのではないかと懸念する報告があります。また、生成AIが高度な音声複製詐欺に利用される可能性を懸念するとの報告もあります。この様な詐欺は、家族や愛する人の声を使って金銭を脅し取るのです。
- **規制されていない。**AIイノベーションの急速なペースは、世界的な規制や法的政策の進展を上回っています。AIの決定に対する異議申し立ての手段が限られています - 「人間と話せませんか？」など...

評価するための標準化されたフレームワークがなければ、主張を検証し、既存の作業に基づいて構築し、責任ある開発を確実にすることが困難になります[20]。現在、ほぼすべての国（付録D）と企業組織（付録E）が、GenAIの開発と展開を管理するためのフレームワークを開発しています。これらのフレームワークはそれぞれ異なる目的を持っていると主張する人もいます。一方で、組織は同じ目的のためにフレームワークを再開発し、組織文化のサポートも行っています。課題は、これらのフレームワークを比較・整理し、誰もが責任を持ってGenAIソリューションを開発・展開する際に参考にできる出発点となる概要をまとめることです。

責任あるGenAIフレームワーク（RGAF）の目的は、GenAIが倫理的かつ公正で、社会全体に有益であり、かつ実証済みで、普遍的に受け入れられる方法で設計、適用、利用されることを促進することです。RGAFは、差別がなく、公正で、持続可能で、プライバシーを尊重し、安全なAIシステムを構築することの重要性を強調しています。また、規制やガイドラインを遵守し、倫理的な意思決定プロセスを採用し、AIシステムが人類全体に利益をもたらすことを保証することも含まれています（付録A：AIの定義）。

AIの進歩の各時代（エキスパート システム、ビッグ データ分析、機械学習からディープラーニング、そして現在の生成AI）は、テクノロジーの機能の大幅な拡張だけでなく、責任ある信用性の高いソリューションを生み出すことの重要性と課題も示してきました[21]。2019年10月、LFAI and DataのTrusted AI Committeeが設立されました。2024年2月、Trusted AI Committeeは新たに設立されたGenerative AI CommonsにおけるResponsible AI Workstreamとなりました。ここ数年、責任あるAIへの注目は、（多くの組織が発表したマニフェストを通じて）業界と研究コミュニティの両方で高まり続けており、従来の機械学習およびディープ ラーニング モデルにおいて堅牢性、公平性、説明可能性に取り組む多くのオープンソース プロジェクトの導入がその推進力となっています。さらに、信頼できる会話型AIに焦点を当てたTrustmark initiativeも最近登場しました。

本研究論文は以下のような構成になっています。以降のセクションでは、9つの主要な視点を持つ責任ある生成AIフレームワークを紹介します。第3セクションでは、EU AI法、NIST AIフレームワーク、シンガポールAI戦略、中国のAI開発計画という4つの主要なグローバル フレームワークを比較対照します。第4セクションと第5セクションでは、RGAFとModel Openness Frameworkの関係を説明し、複合AIシステムのコンテキストでRGAFを使用する方法について簡単に説明します。第6セクションでは、AIシステムが単なるモデルや学習データではなく、複雑適応型サービス システムとして扱う必要がある理由を説明します。最後に、今後の応用研究計画について述べます。

2. 責任ある生成AIフレームワーク

このセクションでは、責任ある生成AIソリューションの設計、開発、展開の文脈における9つの主要な視点について、定義、課題、提案されたソリューションとともに紹介します。

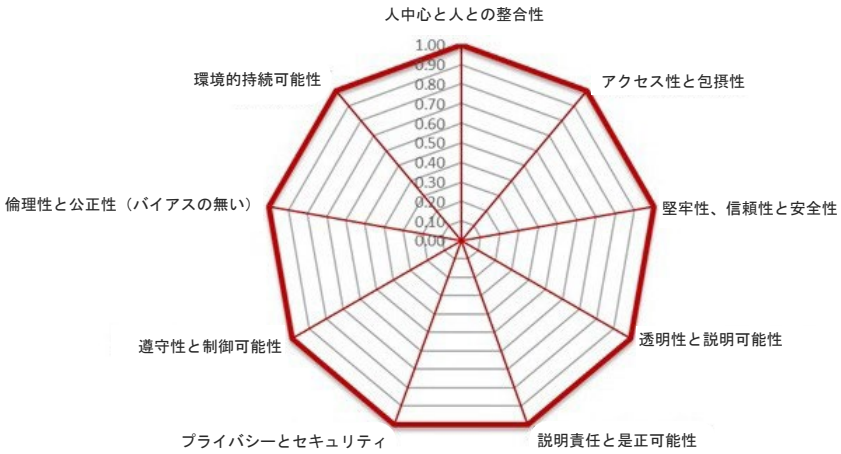


図1：責任ある生成AIフレームワークの視点

2.1. 人中心と人との整合性

定義

- 人中心のAIは、人間の入力、相互作用、協働から学習します。多くのアルゴリズムは、脳理論や社会科学といった人間ベースのシステムから来ています。これらは、人間が関与する相互作用がAIの学習をより良く、より速くするタイプのソリューションです。また、AIシステムは、長期的な価値を実現するために、人間の価値観と社会のニーズを念頭に置いて設計されるべきです[22] [23] [24] [25]。
- **自立性**：AIシステムは人間の自律性を尊重し、自己決定を損なわないようにする必要があります。

課題

- AIシステムが人々の生活を向上させるように設計され、害を及ぼさないことを保証することです。
- AIシステムが多様な人間の価値観を理解し、尊重することを保証することです。
- AIの自律性と人による制御のバランスを取ることです。
- AIが人間の認知バイアスによって悪影響を受けないようにすることです。

改善の可能性

- AIシステムの設計と導入に多様なステークホルダーを関与させます（付録B：AIステークホルダー）。
- 人間の価値観との継続的な整合性を確保するための、堅牢な指標、測定、フィードバック メカニズムを実装します。
- 人中心のAIを設計するためのガイドラインと標準を策定します。

他の視点との相互依存性

- 人中心のAIは、人間のニーズに焦点を当てることで倫理的に問題のある結果につながらないようにするための羅針盤として機能する倫理的なAIを必要とします。
- 倫理的なAIには、人中心のAIが必要です。人々の具体的な経験に基づいてその原則を確立し、それらの原則をシステム設計において実用的で意味のあるものにする役割を果たします[26]。

2.2. アクセス性と包摂性

定義

- AIシステムは、あらゆる個人とグループのニーズと権利を考慮し、社会のあらゆる階層がアクセス可能で、様々な状況や集団において信頼性と一貫性のある性能を提供できる包摂的なものでなければなりません。また、個人、企業、そして国が、法外なコストや社会文化的障壁に直面することなく、AIシステムにアクセスし、恩恵を受けることができる能力も意味します。

課題

- 情報格差や不平等なAIシステムへのアクセスです。
- 能力や経歴に関わらず、誰もがAIにアクセスできることを保証することです。
- AIツール、インフラストラクチャ、熟練した人材の獲得に必要な資金は、中小企業や発展途上国にとって高額になりがちです。これが参入障壁となり、AIの普及を阻害しています[27] [28]。

改善の可能性

- 政府や組織はAIシステムの導入にかかる経済的負担を軽減するために補助金、助成金、低金利融資などを提供することができます。**オープンソース**のAIツールやプラットフォームもコストを削減し、AIへのアクセスと導入を促進することができます[29]。
- ブロードバンド アクセスの拡大やコンピューティング能力の強化といったデジタル インフラストラクチャへの投資は、AI導入に必要なリソースが不足している地域のギャップを埋めることができます[30]。
- 国際的なパートナーシップや連携は、リソースと専門知識の共有を促進し、AIシステムの構築と取得を容易にします。

他の視点との相互依存性

- 倫理性：包摂的AIは、倫理基準と社会的価値観に適合する必要があります。これには、プライバシーの確保、公平性の促進、AIアルゴリズムにおけるバイアスの回避が含まれます。倫理原則を遵守することで、AI開発者はAIシステムの信頼性と責任ある利用に対する信頼を高めることができます。
- 透明性：データやモデル アーキテクチャを含む透明なモデルへのアクセスは、スキルの習得やAIシステム構築の参入障壁を低減します。

2.3. 堅牢性、信頼性と安全性

定義

- 信頼性の高いAIシステムは、困難な環境や予測不可能な環境下においても、意図された機能を一貫して正確、安全、かつ確実に実行します。
- 信頼性の高いAIシステムとは、経験的証拠や定性的な議論に基づいて単に安全であると想定されるのではなく、**安全性**が証明できるシステムです[31]。

課題

- 多様な現実世界の状況において、AIシステムが信頼性と堅牢性を備えて動作することを保証することです。
- 信頼性を阻害する可能性のある技術的制限を克服することです。
- 抜け穴を回避することは容易ではないが、望ましいAIの振る舞いを正式に定義することです[32]。
- 複雑なAIシステムやワールド モデルに対して、大規模な形式的検証を実施するには、高い計算能力必要になることです。

改善の可能性

- 堅牢で、様々な設定にわたって汎用性の高いAIモデルを開発します。
- 自傷行為の可能性を含む有害な状況を認識し、緩和するシステムを設計します。
- コンテンツ フィルターと、このようなリスクを監視する機能を実装します。
- システムが自傷行為、暴力、違法行為などを助長するコンテンツを生成しないように安全対策を講じます。
- 安全でない出力をユーザーに表示する前に検出するためのコンテンツ フィルタリングを実装します。
- 生成された出力の事実の正確さを確認するチェック機能を組み込みます。
- システムが不安定な場合や検索品質が低い場合に、適切な縮退運転を実装します。
- 信頼性を確保するために、AIシステムを継続的に監視および改善します[33]。
- 信頼性検証タスクを管理しやすいサブ問題に分割するなど、AIによる信頼検証を支援します。
- 既存のベンチマーク（MLCommons[34]など）を活用して、モデルに関連する安全性リスクをランク付けします。

他の視点との相互依存性

- 説明可能性：信頼できるAIは、その決定と行動について説明を提供する必要があります。これは、特に医療、教育、自動運転などの重要なアプリケーションにおいて、ユーザーがシステムを理解し、信頼するために重要です[35]。

2.4. 透明性と説明可能性

定義

- AIシステムは、その運用と意思決定において、透明性を保ち、ユーザーや利害関係者が理解しやすい説明を提供する必要があります[36]。
- 透明性の高いAIシステムは、**収集されるデータ、そのデータの使用方法と保存方法、**そして**誰がデータにアクセスできるのか**に関する情報を共有します。また、その目的をユーザーに明確に示します[37]。
- 透明性の向上は、AI利用者にAIモデルやサービスがどのように作成されたかについて深く理解できる情報を提供します。これにより、モデルのユーザーは、それが自身の状況に適しているかどうかを判断できるようになります。
- AIシステムは、技術的な専門知識の有無にかかわらず、*人間*が容易に理解し、信頼できるものである必要があります[38]。

課題

- AIモデルの複雑さ、特に、ディープ ラーニング モデルの複雑さは、解釈を困難にする可能性があります。
- モデルの複雑さと説明可能性の間にはトレードオフがあります。
- 機密情報を損なうことなく透明性を確保することです。用途によって異なりますが、公共財データで学習されたAIモデルは、影響力のある成果につながる可能性があります。公共財データとは、例えば、鳥の移住パターン、落雷、気象データなどに関するデータです。

改善の可能性

- 詳細なドキュメントとライセンスを備えたライフサイクルを構築するAIモデル全体のオープンソース化します。

- 複雑なモデルから説明を抽出するための解釈可能なモデルと技術を開発します。
- AIシステムの透明性を義務付ける規制枠組みを構築します。
- AIによる意思決定の理解を深める視覚化ツールなどのツールを活用します。

2.5. 説明責任と是正可能性

定義

- 信頼を築き、透明性を確保し、規制要件を満たすために不可欠です。
- AIシステムの行動と決定に対する責任と説明責任を明確に割り当てる機構が必要です。
- 説明責任には、3つの主要な項目が含まれます。1）誰が、誰に対して説明責任を負うのか、2）何に対して説明責任を負うのか、3）どのように説明責任を果たし、是正するのか[39]、です。

課題

- 複雑なAIシステムの意思決定プロセスを追跡し、バイアス、倫理的懸念、望ましくない行動を特定することです。
- AIの説明責任に関する法的および倫理的枠組みを確立することです。
- 組織および個人がAI関連の結果に対して説明責任を負うことを保証[40]することです。
- 説明責任の必要性と、開発スピードや費用対効果といった他のビジネス上の優先事項とのバランスを取ることの困難であり、責任あるAIの実践において妥協につながる可能性があります。
- AIシステムの複雑さ、特にGenAIモデルの複雑さは、その動作を理解し解釈することを困難にし、潜在的なバイアスや倫理的懸念を特定して対処する能力を阻害します。
- AIの説明責任を測定・評価するための広く受け入れられた標準やツールが存在しないことから、達成状況を評価し、異なるシステム間でパフォーマンスを比較することが困難になっています。
- AIにおけるイノベーションの促進と、説明責任を確保するために必要な規制の実施との間の適切なバランスを見つけることは、複雑な社会的課題です。

改善の可能性

- AIシステムには、包括的なドキュメント（例えば、モデル カード、データシートなど）と監査証跡（例えば、データ系統の追跡、モデルのバージョンなど）の実装が不可欠です。
- ミスや意図しない悪影響への対応に関する明確なポリシーを策定します。
- 責任ある開発と展開を監督する責任者を任命します。
- これらのポリシーをシステムに実装し、AIアプリケーションを継続的に監視・監査を実施します。
- ユーザーが肯定的または否定的な意見を報告できるフィードバック機構を提供し、それに基づいてシステムを継続的に改善します。
- AIコンテキストにおいて、説明責任を明確に定義したポリシーと規制を策定します。
- AI開発者とユーザーの間で、責任と倫理意識の文化を促進します。
- さまざまなコンテキストに、効果的に一般化できるAIモデルを開発します。

- 不確実性や低品質な検索に対して、適切な縮退運転を実装します。
- 信頼性を体系的に評価するため、AI支援された検証を用います。
- 有害な出力を防ぐための包括的なコンテンツ フィルタリング システムを実装します。
- 自傷行為、暴力、または違法行為を助長するコンテンツの生成を防ぐための安全策を構築します。
- 潜在的に有害な状況を認識し、緩和するシステムを設計します。
- ベンチマーク（例えば、MLCommons）を用いて、モデルの相対的な安全性を評価します。
- 生成されたコンテンツの事実の正確性を検証します。
- 継続的なシステム監視を実施します。
- 性能データに基づいて、反復的な改善を実施します。

他の視点との相互依存性

- 透明性：AIシステムが説明可能で透明性がある場合、規制、倫理、プライバシーの観点からシステムの説明責任を監視および評価することが容易になります。

2.6. プライバシーとセキュリティ

定義

- AIのためのセキュリティ：様々な脅威からデータとシステムのセキュリティを確保し、信頼性と安全性を確保するためのプロセス[41]です。
- AIのためのプライバシー：AIシステム（インテリジェント エージェント[42]を含む）は、個人のプライバシーを保護し、個人データの収集、保管、使用に伴う倫理的および法的影響を軽減する手段を提供する必要があります[43]。
- セキュリティとプライバシーは区別されますが、AIの文脈では重複する部分があります。
- どちらも目的は、データ保護にあります。セキュリティは外部からの脅威に焦点を当て、プライバシーは倫理的かつ責任ある利用に焦点を当てています。
- セキュアなシステムはプライバシーの維持に不可欠であり、プライバシー保護技術はセキュリティを強化することができます。
- どちらも、**責任あるAI**の開発と展開のために、堅牢な規制と倫理ガイドラインを必要とします。

課題

- 機密情報のデータ侵害および不正アクセスのリスクがあります。
- データ保護規制を確実に遵守することです。
- データアクセスの必要性とプライバシーへの懸念のバランスさせることです。
- GenAI固有の脆弱性を悪用する**敵対的プロンプト、ジェイルブレイク、プロンプト インジェクション攻撃**などの、AIシステムに対する敵対的攻撃への対策[44]することです。

改善の可能性

- 強力な暗号化とアクセス制御を実装します。これには、例えば、機密コンピューティング（Confidential Computing）のベスト プラクティスを経由したフォーマット保持暗号化[45]（Format-Preserving Encryption：FPETS）や**完全準同型暗号**[46]（Fully **Homomorphic Encryption**：FHE）が含まれます。
- 学習や検索コーパスで使用される個人データが適切に匿名化されていることを保証します。

- AI開発において**プライバシー・バイ・デザイン**の原則を採用し、これに差分プライバシー、**連合学習（federated learning）**、**分散学習（split learning）**[47]などの技術を組み込み、モデル推論中の攻撃を軽減することが含まれます。
- セキュリティ対策を定期的に更新および監査します。これには、AIシステムの動作の**監視**、**異常の検出**、AIシステム/エージェントに対する新たな脅威に対する**防御の更新**が含まれます。
- **サンドボックス環境**を導入して、ローカルとリモートの両方でAIシステムの機能とリソースへのアクセスを制限し、不正なアクションを防止します。
- AIシステムで使用されるデータの品質、完全性、プライバシーを確保することは説明責任を果たす上で非常に重要ですが、大規模で複雑なデータセットの場合は困難になる可能性があります。
- パラメータ効率の高い微調整（PEFT：Parameter-Efficient Fine-Tuning）[48]を採用し、ユーザーのプライバシーを保護しながらAIシステムをパーソナライズします。ユーザーのチャット履歴を安全に管理し、GenAI外部の更新可能なメモリを活用することで、基盤モデルを損なうことなくシステムのパフォーマンスを向上させます。

- **他の視点との相互依存性**
  - 透明性：理解することはセキュリティの前提条件です。個人は、AIシステムが自身のデータをどのように使用しているかに関する情報にアクセスし、その使用をある程度制御できる必要があります。
  - 倫理性：プライバシーとセキュリティはどちらも倫理的なAIシステムを開発するための鍵であり、同時にAIにおける倫理は、セキュリティとプライバシーの両方を確保するAIシステムを構築するための道徳的な指針を提供します。
  - 説明責任：真正性（Authenticity）はセキュリティの5つの柱[49]の1つであり、**説明責任**を確実に果たすために不可欠です。

2.7. 遵守性と制御可能性

定義

- 遵守するAIシステムは、法的、倫理的、および規制上の基準を遵守します。これには、AIシステムがデータ保護規制、差別禁止法、業界固有の基準など、様々な法律やガイドラインに遵守していることを保証することが含まれます。その目的は、AIシステムによる危害を防ぎ、責任ある倫理的な利用を確保することです[50] [51] [52] [53] [54]。
- 制御可能性：AIシステムを効果的に管理・指揮し、意図したとおりに動作し、必要に応じて修正またはシャットダウンできることを保証する能力を定義します[55]。

課題

- 急速に変化する規制環境に対応することです。
- 必要に応じてAIシステムを人間が制御および手動で止められることを保証することです。
- 人間の監視なしに動作する自律型AIシステムのリスクに対処することです。
- 動的な遵守状況評価の自動化を実装することは、困難を伴い、法規制の進化に合わせて継続的な更新が必要となります[56]。

改善の可能性

- 常に情報を入手し、規制の動向に関与します。
- 設定用取っ手などの制御メカニズムを活用し、組み込み型制御メカニズムを持ったAIシステムを設計します。
- 定常的に遵守状態を監査し、リスク評価を実施します。

- 規則や規制によって強制される制約を遵守しながら、リアルタイムの遵守状況評価を提供する手法を活用します（例えば、コンプライアンス カード アーティファクトや自動分析アルゴリズム[57]）。
- データを使用する前に同意を得ることで、グローバルなプライバシー基準（GDPR、DMA、他）への遵守を確保します。
- ユーザーに自身のデータに対するコントロールを提供し、収集をオプトアウトできるようにします。
- GenAIが特定のアプリケーションに適切かつ有用なテキストを生成するための制御メカニズムを導入します[58]。

他の視点との相互依存性

- プライバシーとセキュリティ：遵守状況をまとめた資料を伴い機密データを扱う際には、プライバシーとセキュリティへの影響を慎重に考慮する必要があります。
- 透明性：遵守は必ずしも透明性を意味するわけではありませんが、透明性は、ルールと規制を同時に遵守する責任あるAIシステムを構築する上で鍵となります。理想的には、この2つは重なり合い、互いに補完し合う必要があります。
- 説明責任と是正可能性：AIシステムは、このフレームワークで期待されるすべての視点について、説明責任を果たし、状況に応じて是正する準備ができている必要があります。

2.8. 倫理と公平性（バイアスの無い）

定義

- 倫理的AIとは、道徳原則と社会的価値観に沿った方法でAIシステムを開発・利用することです。AI倫理の第一の目標は、AIシステムの有益な影響を最大化しつつ、リスクと悪影響を最小限に抑えることです。倫理的AIには、その信用性と信頼性に貢献するいくつかの重要な視点が含まれています[59]。
- 個人または集団への不公平な扱いにつながる可能性のあるバイアスを避けています。
- 善行：AIシステムは、幸福をもたらす様に設計され、人類と環境に利益をもたらすことを促進しています。
- 公正性：AIシステムは、公平性と利益と費用の正当な分配を促進するべきです。
- 説明可能性：AIシステムは透明性が確保され、意思決定プロセスは説明可能であるべきです[60]。

課題

- 差別につながる可能性のあるバイアスを特定し、軽減することです。
- AIの恩恵が社会全体に公平に分配されることを保証する事です。
- 倫理的配慮と実用的な実装のバランスさせることです。
- 説明責任：AIシステムの行動に対する責任の所在を明らかにすることは、特にAIシステムの自律性が高まるにつれて困難になる可能性があります[61]。
- 急速な技術進歩：技術革新のペースが速いため、倫理的枠組みの構築や潜在的なリスクの軽減が困難になっています[62]。
- 測定可能な指標の欠如、すなわち、AIシステムの倫理性をどのように測定すればよいのか[63]、ということです。

改善の可能性

- AIシステムのバイアス監査と公平性評価の実施です（付録C：責任あるAI評価ツールと手法のサンプルリスト）。
- 偏ったコンテンツや不快なコンテンツを最小限に抑えるための、検索に使用する外部知識を慎重にキュレーションおよびフィルタリングします。

- よりバランスの取れた視点を促すなど、GenAIの出力におけるバイアスを検出し、軽減するための技術を実装します。
- アプリケーションが多様なユーザーにサービスを提供し、異なる視点の公平な表現を保証します。
- 多様なトレーニング データセットにおける表現と多様性のある開発チームを確保します。
- 公平性の内訳と法的視点とのバランスの確立について継続的に研究します[64] [65]。
- AI開発プロセス全体を通して、ユーザー、開発者、倫理学者、政策立案者など、多様なステークホルダーを関与させ、公平な代表を確保します。
- 倫理ガイドラインと監督委員会を設置します。
- 設計、開発から展開、監視に至るまで、AIライフサイクル全体にわたって倫理的な配慮を組み込みます[66]。
- データとアルゴリズムの偏りを識別して軽減するための手法を開発し、実装します。
- 多様な視点が考慮されるよう、開発者、倫理学者、政策立案者、そして一般の人々とも連携を推進します。
- 教育とオープンな対話を通じて、AIとその倫理的影響についての一般理解を深めます。
- AIシステムが進化し、新しいデータやコンテキストに適応するにつれて、その倫理的影響を定期的に評価します。
- システムの潜在的な誤用や悪影響について慎重に考慮します。
- 多様な関係者からフィードバックや監視を受けます。

他の視点との相互依存性

- 説明責任があり、是正可能であることです。
- 透明性があり、説明可能であることです。
- アクセスしやすく、包摂的であることです。

2.9. 環境的持続可能性

定義

- AIの開発と展開は環境的に持続可能であり、環境への悪影響を最小限に抑える必要があります[67]。

課題

- AIのトレーニングと導入に伴う高いエネルギー消費量と二酸化炭素排出量です。
- データセンター、AIファクトリー、コンピューティング インフラストラクチャの環境への影響です。
- 技術進歩と環境持続可能性のバランスです。
- 具体化されたAIと運用AIの両方のカーボン フットプリントを維持・監視します（比率は2：1）[68]。
- 効率性の向上は単なる副次的な要素に過ぎず、リバウンド効果（効率性の向上が利用の増加につながり、二酸化炭素排出量の増加につながる）[69]を完全に解決するには不十分です。

改善の可能性

- データセンターの電力供給と大規模言語モデルの学習のために、再生可能エネルギー源への投資を優先します。
- よりエネルギー効率の高いAIアルゴリズムとハードウェアを開発します。

- データセンター向け再生可能エネルギー源に投資します。
- AIシステムの二酸化炭素排出量を削減する最優良事例を推進します。
- 継続学習（CL：Continual Learning）[70]などの手法を採用することで、段階的に学習するため、データセット全体の再学習を必要とする従来のAIモデルと比較して、必要な計算リソースが少なく、効率性と拡張性が向上します。
- 教師ありファイン チューニング（full supervised fine tuning）の代替として、パラメータ効率的なファイン チューニング（PEFT：Parameter Efficient Fine Tuning）を使用します。
- AIライフサイクル全体の持続可能性を評価します[71]。
- AIライフサイクル全体にわたるシステムの二酸化炭素排出量を測定するための遠隔測定と可観測性を導入します。
- 耐久性、修理性、モジュール性を考慮した設計により、既存のハードウェアを最大限に活用し、頻繁な交換の必要性を軽減します（ハードウェア寿命の延長と、排出される二酸化炭素排出量の削減に貢献します）。
- 半導体製造における二酸化炭素排出量を最小限に抑えるための研究開発への投資と、代替材料の探索を行います。
- システムのトレーニングと運用にかかる環境コストを公開します。
- 適切なソリューションと実行されたアクションを選択する際に、持続可能性スコアを評価します。
- 雇用創出、データ プライバシー、ヘルスケア、環境・社会への影響など、AIの基礎フレームワークが整備されていない発展途上国で、非倫理的なテストが行われないようにします。
- 大規模言語モデルの学習では、再生可能エネルギー源と効率的なハードウェアの使用を優先します。システムの二酸化炭素排出量について透明性を確保し、長期的に最小限に抑えるよう努めます。
- 環境負荷を軽減するためのカーボンオフセット計画を公開します。

他の視点との相互依存性

- 人中心と人との整合性
- 堅牢性、信頼性と安全性
- 説明責任と是正

次のセクションでは、EU AI法のフレームワーク、NIST AIフレームワーク、シンガポールAI戦略、そして中国のAI開発計画がカバーしている主要な分野について考察します。RGAFは、研究開発とニーズに加えて、これらすべての分野を網羅していることがお分かりいただけるでしょう。

3. 責任あるAIの視点をフレームワーク間で比較対照

このセクションの目的は、さまざまなAIフレームワークの概要を示し、提案したRGAFフレームワークがグローバルな視点で利用可能なフレームワークの要件をどのように満たすかを説明することです。

表1：4つの主要なグローバルフレームワークの概要

	EU AI法フレームワーク[72]	NIST AIフレームワーク[73]	シンガポールAI戦略[74]	中国AI開発計画[75]
人中心と人との整合性	欧州委員会により選定されたAIハイレベル専門家グループ（AI HLEG：AI High-Level Expert Group on Artificial Intelligence）のガイドラインで、AIは人間の代理として認識されています。AIシステムは人間に役立ち、人間の尊厳と個人の自律性を尊重し、人間によって適切に制御および監視できるツールとして開発および利用される必要があります。	リスク管理フレームワーク（RMF：Risk Management Framework）では、人間とAIの構成と監視、人中心の設計、評価、テストとして認識されています。	システムの使いやすさ、倫理原則との整合性、システムが人間の能力を拡張する可能性などを文書化しています。	中国政府は、人間中心で人間の価値観に沿ったAIの開発の重要性を強調しているが、明確なガイドラインはありません[76]。
アクセス性と包摂性	加盟国による社会的に有益な成果を支援するAIソリューションの研究開発を奨励しています。社会福祉の重要性とともに、データ主体の権利と自由に対する高いリスクを特定するためのメカニズムの必要性を強調しています。	技術的要因と社会的要因の複雑な相互作用、および人々、組織、エコシステムへの潜在的な損害を考慮する社会技術的アプローチを採用しています。	システムと社会的価値観の整合性、およびシステムが意図しない結果を生み出す可能性を文書化しています。	中国政府はAIが社会や環境に及ぼす潜在的な影響を認識しており、これらの要素を考慮した責任あるAI開発を求めているが、明確なガイドラインはありません[77]。

	EU AI法フレームワーク	NIST AIフレームワーク	シンガポールAI戦略	中国AI開発計画
堅牢性、信頼性と安全性	技術的な堅牢性は、高リスクAIシステムの重要な要件です。高リスクAIシステムの正確さと堅牢性に対処するためのベンチマークと測定手法の開発において、関係者の協力が必要であることが、文書全体を通して言及されています。	RMFでは信用できるAIシステムを特性として定義されており、ガイダンスを提供しています。堅牢性には、AIシステムが想定される用途で機能するだけでなく、予期しない環境で動作した場合に人への潜在的な危害を最小限に抑えることも求めています。	冗長性やエラー訂正技術などを使用して、正確性とエラーや障害に対する堅牢性を考慮して設計します。	AIシステムなどのソフトウェア アーキテクチャの最良事例を適用することは、さまざまな条件下で安定的かつ一貫して動作し、一部のコンポーネントに障害が発生した場合でも動作を継続できることです。
透明性と説明可能性	AI HLEGの拘束力のないAI倫理原則として規定されています。AIシステムの開発と使用は、追跡可能性と説明可能性を実現し、人間がAIシステムと通信または相互作用していることを認識できるようにします。	RMFにおいて信用できるAIシステムの特性として定義されており、ガイダンスを提供します。透明性は、AIライフサイクルの段階に基づき、AIアクターの役割や知識、あるいはAIシステムと相互作用する人々に合わせて調整された適切なレベルの情報へのアクセスを提供します。説明可能性とは、AIシステムの動作の基盤となるメカニズムの表現を指します。	使用されるアルゴリズムの解釈可能性、ドキュメントの可用性、およびシステムがユーザーに有意義なフィードバックを提供する能力などの意思決定の明確で理解しやすい説明を文書化します。	サービスタイプの特性に基づいて、組織は生成AIサービスの透明性を高め、生成されたコンテンツの正確性と信頼性を向上させるための効果的な対策を講じるべきと推奨しています[78]。

	EU AI法フレームワーク	NIST AIフレームワーク	シンガポールAI戦略	中国AI開発計画
説明責任と是正可能性	AI HLEGの拘束力のないAI倫理原則として規定されています。	RMFでは、信用できるAIシステムの特性として定義されており、ガイダンスを提供しています。説明責任は透明性を前提としています。透明性とは、AIシステムと相互作用する個人が、その相互作用を行っているかどうかに関わらず、AIシステムの情報と出力がどの程度利用可能であるかということです。	責任と賠償責任を確保するためのメカニズム、すなわち監査証跡の可用性、意思決定プロセスの透明性、明確なガバナンス構造の存在など、文書化しています。	AI関連インシデントに対する賠償責任メカニズムの確立を含む、AIシステムの説明責任に関するガイドラインを示しています[79]。
プライバシーとセキュリティ	AI HLEGの拘束力のないAI倫理原則として規定されています。AIシステムは、プライバシーとデータ保護のルールに従って開発・使用され、品質と完全性に関して高い基準を満たすデータを処理します。	RMFでは、信用できるAIシステムの特性として定義されており、ガイダンスを提供しています。プライバシーとは、人間の自律性、アイデンティティ、尊厳を守るための規範と実践を指します。	プライバシーとセキュリティの保護手段、すなわち個人データの保護、暗号化やその他のセキュリティ対策の使用、攻撃に対するシステムの堅牢性など、を文書化しています。	サイバーセキュリティ法[80]や個人情報保護法[81]など、データのプライバシーとセキュリティに関する厳格な規制を示しています。

	EU AI法フレームワーク	NIST AIフレームワーク	シンガポールAI戦略	中国AI開発計画
遵守性と制御可能性	この法律は、AIシステムを潜在的なリスク レベルに基づいて分類しており、「許容できないリスク」のシステムは禁止され、「高リスク」のシステムには広範なコンプライアンス対策が必要となり、「低リスク」のシステムにはそれほど厳しくない要件が課せられています。	遵守性と制御可能性のリスクにより、AIは組織にとっても社会にとっても導入・活用が非常にユニークで挑戦的な技術になっています。適切な管理がなければ、AIシステムは個人やコミュニティにとって不公平または望ましくない結果を増幅、永続化、あるいは悪化させる可能性があります[82]。	シンガポールAI戦略は、「モデルAIガバナンス フレームワーク（Model AI Governance Framework）」に概説されている倫理ガイドラインと自主基準の枠組みを通じて、責任あるAIの開発と展開を促進することで、「遵守」を重視しています[83]。	中国政府は最近、遵守体制の構築と実施を推奨し、関連法を制定しました。遵守問題の不適切な処理は、企業に損害を与えるだけでなく、関係者が個人的な責任を問われる可能性があります。
倫理と公平性	倫理的かつ公正であることは、AI HLEGにおける拘束力のないAI倫理原則として規定されています。AIシステムの開発と利用において、EU法または国内法で禁止されている差別的影響や不公平なバイアスを回避することが重要です[84]。	RMFでは、信用できるAIシステムの特性として定義されており、ガイダンスを提供しています。公正さは文化によって異なり、用途によっても変化する可能性があります[85]。	トレーニング データの多様性、使用されるアルゴリズムの公平性、および意図しない結果の可能性を文書化しています。	顔認識技術などのAIアプリケーションにおける差別を防止するためのガイドラインを示しています[86]。

	EU AI法フレームワーク	NIST AIフレームワーク	シンガポールAI戦略	中国AI開発計画
環境的持続可能性	AIシステムが環境の持続可能性に与える影響を評価し、最小限に抑えることは重要です。AIシステムの効率的な設計、トレーニング、運用のためのエネルギー効率の高いプログラミングと技術、そして使用される計算リソースとエネルギー消費量を文書化することが義務付けられています。	AIシステムが環境の持続可能性に与える影響を評価し、最小限に抑えることは重要です。AIシステムの効率的な設計、トレーニング、運用のためのエネルギー効率の高いプログラミングと技術、そして使用される計算リソースとエネルギー消費量を文書化することが義務付けられています。	再生可能エネルギー源を使用し、電子廃棄物を削減し、資源を最小限に抑えるグリーンAIの原則を示しています。	グリーンAIの原則は、再生可能エネルギー源の利用、電子廃棄物の削減、資源の最小化を示しています。AIが環境に与える影響に関する技術的枠組み、方法、指標を標準化し、産業発展と環境保護のバランスをとることが必要です[87]。

RGAFは、上記のすべてのガイドライン、要件、ポリシーを網羅するグローバル フレームワークです。オープンソース モデルの開発と展開は、GenAIソリューションの成功の鍵となります。信頼性はGenAIのソリューション導入において、最も重要な要素の一つであり、オープンソース モデルは透明性、説明可能性、そしてその他多くのRGAFの要素を実現することで、信頼性を獲得・構築します（付録D：責任あるAIフレームワークとガイドラインを策定した国と組織のサンプル リスト、付録E：責任あるAIフレームワークとガイドラインを策定した民間企業のサンプル リスト、付録F：関連する責任あるAI活動と組織）。

4. Model Openness Frameworkとの関係

The Model Openness Framework（MOF）[89]は、機械学習およびAIモデルのオープン性と完全性を評価・分類するためのシステムです[89]。このフレームワークは、「オープン ウォッシング」[90]に対抗することを目的として、モデル開発ライフサイクルに関わるすべての成果物の公開を奨励することで責任あるAI開発を促進しており、「オープン ウォッシング」に対抗する上で重要なポイントとなっています。以下はMOFの主な特徴です。これらの特徴をここに記載するのは、RGAFとMOFという2つの補完的なフレームワークを分かりやすく説明するためです。

- 3段階の分類システム（クラスI、II、III）：上位クラスはより完全でオープンなモデルを示し、クラスIはオープン サイエンスの原則に沿ったゴールド スタンダードを表します。
- 完全性のためのコンポーネント：モデル開発ライフサイクル全体にわたるデータ、コード、ドキュメントを網羅します。
- 特定のオープン ライセンス：各コンポーネントに適したオープン ライセンスを推奨します（例えば、コードにはOSI承認ライセンス、データセットにはオープン データ ライセンス）。
- Model Openness Tool：MOF基準に照らしたモデルを評価のためのリファレンス実装を提供し、適格モデルにはバッジを生成します。

表2：MOFフレームワークがオープンRGAを実現する方法

MOF要素	RGAF視点	関係
データセット	倫理性と公正性	データセットを公開することで、データに潜むバイアスを検証し、AIにおける公平性と包摂性を促進します。
データ前処理コード	透明性と説明可能性	モデル トレーニング用のデータ準備手順を公開することで、オープンな前処理コードは信頼と理解を促進します。
モデル アーキテクチャ	透明性と説明可能性	モデル アーキテクチャを共有することで、AIシステムの分かりやすさが向上し、説明可能性に貢献し、分析が可能になります。
モデル パラメータ	アクセス性と包摂性	オープンなモデル パラメータにより、他の人が利用、検証、モデルの挙動を明らかにし再現できるようになり、より広範なアクセスと信頼性の確保につながります。
トレーニング コード	透明性と説明可能性	トレーニングコードを公開することで、モデル開発プロセスを理解し、バイアスや不公平の潜在的な原因を特定できるようになります。

MOF要素	RGAF視点	関係
推論コード	アクセス性と包摂性	推論コードを共有することで、モデルの機能へのアクセスが広がり、性能の独立した検証を可能にします。
評価結果	説明責任と是正可能性	評価結果の透明性は責任を示し、モデルの能力と限界の独立した評価を可能にします。
評価コード	説明責任と是正可能性	評価コードを公開することで、モデルの評価に使用された方法を精査できるようになり、提示性能における厳密さと説明責任が確保されます。
モデル カード	透明性と説明可能性	モデル カードには、モデル、その使用目的、制限、および潜在的な倫理的考慮事項に関する必要不可欠な情報が記載されており、透明性が高まり、責任ある使用が促進されます。
データ カード	倫理と公平性	トレーニング データに関する情報を開示することで、データ カードは、潜在的なバイアスを特定し、公平で包摂的なAI開発を促進するのに役立ちます。
研究論文	透明性と説明可能性	公開されている研究論文では、モデルの開発、理論的根拠、性能について詳細な説明が提供され、AIシステムの包括的な理解が促進されます。
技術レポート	透明性と説明可能性	技術レポートでは、モデルの機能と想定される用途について分かりやすく説明されており、より幅広いユーザーがAIシステムを理解しやすくなります。
サンプル モデル出力	倫理と公平性	モデル出力の例を共有することで、潜在的なバイアス、意図しない結果、さまざまなシナリオにわたるモデルの全体的な動作を評価できます。
MOF設定ファイル	遵守性と制御可能性	構成ファイルはオープン標準の遵守を保証し、リリースされたコンポーネントを構造化された方法で文書化する方法を提供するため、他のユーザーがモデルを評価して制御しやすくなります。
サポート ライブラリとツール	アクセス性と包摂性	モデル開発で使用されるツールを共有することで、開発者は他のユーザーがAIシステムを簡単に使用、適応、構築できるようにし、より広いアクセシビリティと堅牢な共同開発を促進します。

MOF要素	RGAF視点	関係
モデル メタデータ	透明性と説明可能性	オープン メタデータは、モデルの開発、トレーニング データ、および潜在的な制限に関するコンテキストを提供し、透明性を高め、ユーザーがモデルの機能と制約を理解するのに役立ちます。
評価データ	説明責任と是正可能性	オープンな評価データにより、特定のタスクにおけるモデルの性能を独立して検証できるようになり、説明責任が強化され、提示された結果の妥当性が保証されます。

RGAFとは何か、とMOFを使用したオープンRGAFについて理解した後は、GenAIソリューションを開発および展開する方法について見ていきます。

5. デコーディング トラスト フレームワークがRGAFの信頼を実現する方法

前のセクションでは、Model Openness Framework（MOF）が提案しているがRGAFフレームワークにどのように対応しているかについて説明しました。MOFはモデルのオープン性を測定および評価するための有益なツールとして機能します。モデルのオープン性は、責任ある信用できるAIの重要な柱ですが、それが自動的に信頼性や責任につながるわけではありません。オープン モデルの信用性は、（構築方法だけでなく）開発、展開、管理、使用方法など、さまざまな要因に依存します[91]。モデルのオープン性に加えて、**デコーディング トラスト（Decoding Trust）フレームワーク**[92] [93]は、モデルの信用性を定量化しようとします。これは、GPTを含む一連のモデルに焦点を当て、大規模言語モデル（LLM）の*信用性*を評価するための包括的なベンチマークを提供します[94]。

表3：RGAFで信頼を実現するためにデコーディング トラスト フレームワークを活用する方法

信用性の観点 (DecodingTrust)	**責任ある生成AIの視点**	**注記**
有害性	倫理と公平性、アクセス性と包摂性、人中心と人との整合性	有害性は人間の幸福に直接影響を与え、有害な社会的偏見を永続させる可能性があります。
ステレオタイプバイアス	倫理と公平性、アクセス性と包摂性、人中心と人との整合性	ステレオタイプ化は倫理原則に違反し、包摂性を損ないます。
敵対的堅牢性	堅牢性、信頼性と安全性、遵守性と制御可能性、プライバシーとセキュリティ	信頼性の低いモデルは誰もが利用できるわけではなく、悪意のある目的に利用される可能性があります。
分布外堅牢性	堅牢性と信頼性、遵守性と制御可能性	信頼性と制御性には、多様な状況における一貫した性能が不可欠です。
敵対的デモンストレーションに対する堅牢性	堅牢性と信頼性、遵守性と制御可能性	操作されたデモンストレーションは、信頼性が低く制御不能なAIにつながる可能性があります。
プライバシー	プライバシーとセキュリティ	個人データの保護は、責任あるAIにとって最も重要です。
機械倫理	倫理と公平性	AIシステムが人の倫理的価値観と整合していることを保証することは、責任ある利用にとって不可欠です。
公平性	倫理と公平性	公正なAIシステムは、すべての個人とグループを公平に扱います。

6. AIシステムはモデルだけではない

生成AIシステムは、モデルだけでなく、複数のコンポーネントで構成されています[95]。これらのコンポーネントには、モデル、Retriever（関連情報検索）、エージェント、または外部ツール（例えば、関数呼び出しなど）への複数の呼び出しが含まれています[96]。複合AIシステムの一般的な形態の1つは、検索拡張生成（RAG：Retrieval augmented generation）アプリケーションです（例えば、DataBricksの調査によると、LLMアプリケーションの60%が何らかの形のRAGを使用しています）[97]。もう1つの増加傾向にあるパターンは、*エージェント型*または*マルチエージェント型*AIアプリケーションです[98]。これらのAIシステムは、少なくとも1つのモデル、データ検索システム、および追加のツール（例えば、Web検索サービス）を組み合わせます。複合AIシステムは、さまざまなAIモデル、ツール、パイプラインの利点と有用性を活用し、個々のモデルのみを使用する場合と比較して、性能、汎用性、再利用性などのユーザー定義された制約を最適化します。個々のモデルだけでは、エンドツーエンドのユーザー体験と製品全体を生み出すには不十分な場合があります[99]。AIシステムは様々な相互接続された要素から構成されるため、信用性の包括的な評価は、モデルだけでなくシステムのすべてのコンポーネントを網羅する必要があります。次のセクションでは、具体的な例を用いて、提案されているRGAFフレームワークをAIシステムとしてのRAGアプリケーションにどのようにマッピングできるかを示します。同様に、RGAFは、エージェント型AI、マルチエージェント型AIなどの他のパターンにもマッピングできます。

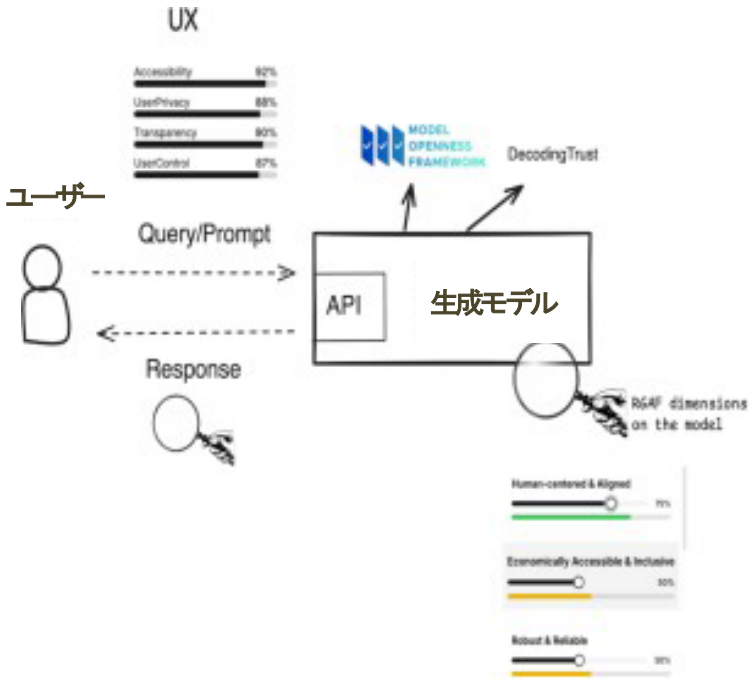


図2：事例研究例：責任あるRAGアプリケーションの構築

シンプルなAIアプリケーションでは、モデルは主要な構成要素の一つとみなされます。モデルに加えて、ユーザーが入力するクエリ/プロンプト、そしてモデルによって生成される応答も考慮します。*クエリ/プロンプト*については、以下に、RGAFの関連する視点をいくつか示します。

1. **人中心と人との整合性**：プロンプト インターフェースはユーザーを念頭に置いて設計し、理解しやすく使いやすいものにする必要があります。AIシステムの本来の目的に沿ったプロンプトを作成するようユーザーを導き、望ましい回答を引き出す必要があります。
2. **アクセス性と包摂性**：プロンプト インターフェースは、障害のあるユーザーや技術レベルの異なるユーザーなど、多様なユーザーに対応する必要があります。これには、複数の入力方法（例えば、音声、テキストなど）のサポート、多言語サポートの提供、明確な指示と実例の提供などが含まれます。

3. **堅牢性、信頼性と安全性**：システムは、曖昧、不完全、またはエラーを含むものも含め、幅広いユーザー プロンプトに対応できる必要があります。予期しない入力を適切に処理し、ユーザーに有益なフィードバックを提供する必要があります。

4. **透明性と説明可能性**：ユーザーは、プロンプトがどのように処理され、モデルの応答にどのような影響を与えるかについて基本的な理解を持つ必要があります。これには、適切なプロンプトとは何かについての明確なガイドラインの提供、効果的なプロンプト例の提供、システムの制限事項の説明などが含まれます。

5. **説明責任と是正可能性**：ログ記録フレームワークとユーザー アクティビティ追跡ツールを使用して、すべてのユーザー プロンプトと関連メタデータを記録します。プロンプト ログを分析して潜在的な問題や不正使用を特定し、システムの説明責任能力を確保します。

6. **プライバシーとセキュリティ**：入力検証ライブラリを用いてユーザー入力を検証、消毒し、悪意のあるコードの侵入を防ぎます。専用ツールを用いて機密情報を匿名化し、アクセス制御を実装してプロンプト データへのアクセスを制御します。

7. **遵守性と制御可能性**：コンテンツ フィルタリングAPIとポリシー適用エンジンを活用し、不適切なプロンプトを検出し、ブロックします。ルールベースのシステムを実装してコンテンツ ポリシーを適用し、問題のあるプロンプトに関するレポート機構を提供します。

8. **倫理性と公正性（バイアスの無い）**：AI Fairness 360などのツールを使用して、トレーニング データに潜むバイアスを分析します。公平性を考慮したNLPライブラリを活用することで、プロンプト処理におけるバイアスを軽減し、公平な対応を確保します。システムのバイアス監査を定期的に実施します。

9. **環境的持続可能性**：エネルギー監視ツールとリソース最適化ライブラリを用いて、プロンプト処理アルゴリズムを最適化し、エネルギー効率を向上させます。クラウド コンピューティング リソースを効率的に活用し、パイプラインのエネルギー消費を監視します。

*応答*について、関連する視点は次のとおりです。

1. **人中心と人との整合性**：関連性があり、有益で、偏見のない応答を生成します。

2. **アクセス性と包摂性**：複数の形式で回答を提供します（例えば、音声合成エンジンによる音声、専用ジェネレーターによる代替形式など）。多言語サポートを提供し、障害のあるユーザーが認識して理解できる回答を提供します。

3. **堅牢性、信頼性と安全性**：専用ツールと自動テスト フレームワークを用いて、関連指標（精度、適合率、再現率）に基づいてモデルの性能を評価します。単体テストと統合テストを実装し、アンサンブル法などの手法を用いて堅牢性を向上させます。

4. **透明性と説明可能性**：LIMEやSHAPなどのモデル説明可能性ライブラリを活用して、応答に影響を与える要因の説明を提供します。専用ツールを使用してモデルの注意を視覚化し、意思決定プロセスに関する洞察を提供します。

5. **説明責任と是正可能性**：ログ フレームワークとモデル監視ツールを使用して、生成されたすべての応答と関連メタデータをログに記録します。モデルのパフォーマンス指標を追跡し、応答ログを分析して潜在的な問題やバイアスを特定します。

6. **プライバシーとセキュリティ**： 専用ライブラリを使用してモデル出力を消毒し、悪意のあるコードの侵入やデータ漏洩を防止します。差分プライバシーなどの技術を実装し、脆弱性を特定するための専用ツールを用いた定期的なセキュリティ監査を実施します。

7. **遵守性と制御可能性**：専用ライブラリを使用してモデル出力を消毒し、悪意のあるコードの侵入やデータ漏洩を防止します。差分プライバシーなどの技術を実装し、脆弱性を特定するための専用ツールを用いた定期的なセキュリティ監査を実施します。

8. **倫理性と公正性（バイアスの無い）**：専門ツールを用いて、公平性指標（機会均等、人口統計的平等）に基づくモデルの性能を評価します。潜在的なバイアスがないか回答を分析し、公平性を考慮したトレーニング手法を用いてバイアスを軽減します。

9. **環境的持続可能性**： エネルギー監視ツールとモデル最適化ライブラリを用いて、レスポンス生成アルゴリズムを最適化し、エネルギー効率を高めます。モデルの圧縮やパラメータを削減するプルーニングといった手法を用いて、パイプラインのエネルギー消費を監視します。

*モデル*については、Model Openness Framework（MOF）やデコーディング トラスト フレームワークなどの既存のフレームワークに準拠することに加えて、RGAFフレームワークをモデルに適用する方法を次に示します。

1. **人中心と人との整合性**：価値に配慮したデザイン フレームワークと参加型デザイン手法を用いて、アーティスト、作家、倫理学者などのステークホルダーをモデル設計プロセスに参画させます。生成AIが人間の創造性と文化的表現に及ぼす潜在的な影響に焦点を当て、モデルの開発と展開に関する明確な倫理ガイドラインを定義します。定期的に倫理的影響評価を実施します。

2. **アクセス性と包摂性**：トレーニングデータが多様な創造スタイルと文化的視点を反映していることを確認します。データ拡張技術と公平性を考慮したトレーニング ライブラリを活用することで、データの不均衡に対処し、バイアスを軽減し、生成されるコンテンツの包括性を促進します[100]。

3. **堅牢性、信頼性と安全性**：大規模かつ多様なコンテンツ データセットおいてモデルを学習させ、データの品質と代表性を確保します。流暢さ、一貫性、独創性、多様性に焦点を当て、様々な創造タスクと指標を用いてモデルを評価します。敵対的トレーニングやモデル アンサンブルといった手法を実装することで、堅牢性を向上させ、望ましくない出力を防止します。

4. **透明性と説明可能性**：アテンション メカニズムを備えたトランスフォーマー モデルなど、ある程度の解釈可能性を備えたモデル アーキテクチャを選択します。モデルの説明可能性ライブラリと可視化ツールを活用し、モデルの創造プロセスと意思決定に関する洞察を提供します。生成モデルの潜在空間と内部表現を理解するための手法を探求します。

5. **説明責任と是正可能性**：モデル監視ツールとバージョン管理システムを用いて、モデルのトレーニング データ、ハイパーパラメータ、パフォーマンス指標を追跡します。製品内でモデルの動作を監視し、生成されたコンテンツの潜在的なバイアスや意図しない結果に細心の注意を払います。モデルの変更を追跡し、必要に応じてロールバックできるようにバージョン管理を実装します。

6. **プライバシーとセキュリティ**：セキュリティ ツールと成功事例を活用し、モデルの重み、コード、トレーニング データを不正アクセスから保護します。連合学習や差分プライバシーなどの技術を活用して、分散データでモデルをトレーニングし、トレーニング データと生成されたコンテンツに含まれる機密情報を保護します。

7. **遵守性と制御可能性**：遵守監視ツールとモデル ガバナンス フレームワークを用いて、モデルの開発と展開が関連規制（GDPR、著作権法など）に準拠していることを確認します。モデルのライフサイクルを管理し、定期的な遵守監査を実施するためのガバナンス フレームワークを実装します。生成される出力のスタイル、コンテンツ、フォーマットを制御するためのメカニズムを開発します。

8. **倫理性と公正性（バイアスの無い）**：バイアス検出ツールと公平性を考慮した評価指標を用いて、トレーニング データとモデル生成コンテンツのバイアス監査を定期的に実施します。公平性を考慮したトレーニング手法とバイアス軽減戦略を採用し、公平で包括的なコンテンツ生成を促進します。有害なステレオタイプの生成や文化的遺物の流用など、モデルの導入による意図しない結果を検出し、対処するための機構を実装します。
9. **環境的持続可能性**：エネルギー監視ツールとモデル最適化ライブラリを用いて、エネルギー効率の高いモデル アーキテクチャとトレーニング アルゴリズムを選択します。計算コストを削減ために、モデルのサイズと複雑さを最適化します。クラウド コンピューティング リソースを効率的に活用し、大規模な生成AIモデルのトレーニングと展開に伴う環境への影響を軽減する手法を検討します。

以下は、追加のコンポーネントを含む別の例です。例えば、Retrieverの場合、最初に使用したインデックス作成プロセス中にトークン埋め込みの結果を保存するためのデータストアが必要です。

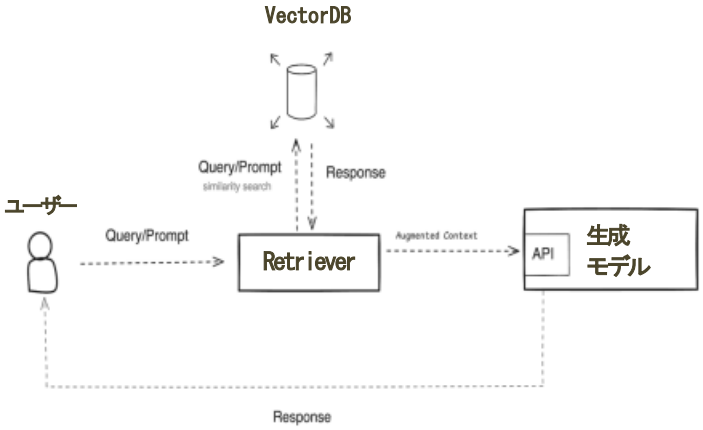


図3：Retrieverコンポーネントを備えたRAGアプリケーション

ユーザーがモデルに送るクエリ/プロンプトに加え、ユーザーのクエリに最も関連性の高い情報チャンクを見つけるためにRetrieverコンポーネントに開始させるクエリ/リクエストも、提案しているRGAFフレームワークを使用して検査および評価する必要があります。Retriever自体とVector DBも、RGAFを使用して設計、検査、評価する必要があります。以下は、サンプルAIシステムのアーキテクチャの各段階と各コンポーネントにおいて、RGAFの各視点をどのように考慮できるかを強調するための追加の考慮事項です。

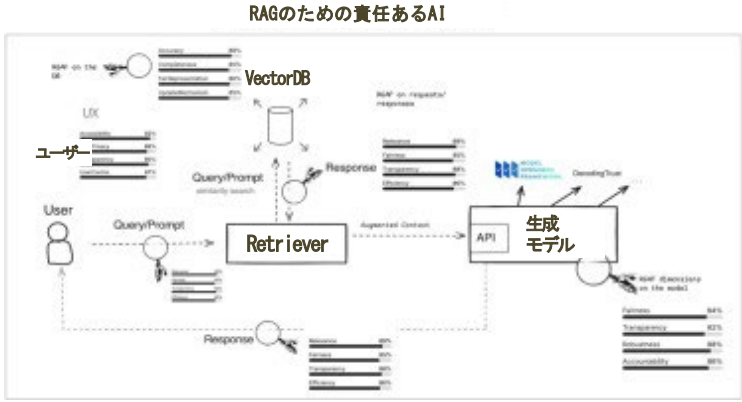


図4：RGAFを使用したサンプルRAGアプリケーション

ユーザー体験（UX）

1. **倫理性と公正性**：データの悪用を防ぐための厳格なガイドラインを強調し、コミュニティのリーダーと協力して、その運営が公衆のプライバシーと権利を尊重するようにします。
2. **環境的持続可能性**：フットプリントを軽減するためのカーボン オフセット計画を公開します。

リクエストと応答

シンプルなAIアプリケーションと同様に、各段階でのリクエスト（クエリ＋迅速な処理）/応答を検査する必要があります。

1. **人中心と人との整合性**：（例えば、ユーザー調査やフィードバックを通じて）リクエストは直感的で、ユーザーのニーズを反映するものであるべきです。一方、クエリは、（例えば、セマンティック検索を使って）関連するコンテキストを検索するものであるべきです。レスポンスは、（例えば、ユーザーからのフィードバックや評価を通じて）ユーザーの期待に沿った有益な情報であり、（例えば、明確な言葉遣いや視覚化を使って）明確に提示されるものであるべきです。
2. **アクセス性と包摂性**：リクエスト インターフェースとレスポンスは、多様なユーザーに対応する必要があります。クエリは、（例えば、多言語マルチモーダル検索を使って）多様な情報源にアクセスでき、レスポンスは、（例えば、音声合成や代替出力形式を提供して）アクセシブルな形式で提示される必要があります。
3. **堅牢性、信頼性と安全性**：システムは、（例えば、ファジー マッチングや自然言語理解を使って）曖昧なものも含め、多様なリクエストやクエリを処理できる必要があります。クエリは、（例えば、堅牢な検索アルゴリズムを使って）ノイズの多いデータにも耐性を持つ必要があります。応答は、（例えば、厳格なテストと検証を通じて）正確で、一貫性があり、エラーがないものである必要があります。
4. **透明性と説明可能性**：ユーザーは、（例えば、明確な指示や例の提示により）リクエストがどのように処理されるかと、（例えば、検索プロセスの視覚化により）クエリがどのように情報を検索するかを理解する必要があります。統合プロセスとレスポンス生成は、（例えば、検索した情報ソースの強調表示により）透明性が保たれ、適切な場合には、（例えば、モデル説明可能性技術を使って）説明が提供されるべきです。
5. **説明責任と是正可能性**：リクエスト、クエリ、レスポンスは、（例えば、ログ記録フレームワークや監視ツールを使って）監査およびパフォーマンス分析のために、ログに記録し、追跡する必要があります。各コンポーネントの運用については、（例えば、ドキュメントの作成や役割の割り当てを通じて）明確な責任体制を確立する必要があります。
6. **プライバシーとセキュリティ**：リクエスト、ユーザー データ、クエリ、応答は、（暗号化とアクセス制御を使って）不正なアクセスや変更から保護される必要があります。機密情報は、（例えば、データ匿名化技術を使って）必要に応じて匿名化または非識別化する必要があります。
7. **遵守性と制御可能性**：リクエスト、クエリ、レスポンスは、関連する規制およびポリシーに準拠する必要があります（例えば、GDPRや投稿監視ガイドライン）。（例えば、コンテンツ フィルタリングAPIを使って）不適切なコンテンツをフィルタリングする機構と、（例えば、カスタマイズ可能な設定を通じて）ユーザーがシステムの動作を制御できる機構をする必要があります。
8. **倫理性と公正性（バイアスの無い）**：システムは、（例えば、バイアス検出ツールや公平性を考慮したアルゴリズムの使用によって）リクエストの処理、クエリの作成、応答の生成においてバイアスを回避する必要があります。公平性を考慮した技術は、（例えば、多様なトレーニング データを確保し、公平性指標を評価することによって）システム全体にわたって採用する必要があります。
9. **環境的持続可能性**：リクエスト処理、クエリ実行、応答生成は、（例えば、エネルギー効率の高いアルゴリズムとハードウェアを使って）エネルギー効率を考慮して最適化し、（例えば、再生可能エネルギー源を使って）システムの環境への影響を最小限に抑える必要があります。

Retriever

1. **人中心と人との整合性**：関連性があり、正確で、ユーザーの情報ニーズと価値観に合った情報の検索を優先するようにRetrieverを設計します。
2. **アクセス性と包摂性**：Retrieverがさまざまなソースや形式の情報にアクセスして処理できるようにし、さまざまな視点や知識領域をシステムに組み込むようにします。
3. **堅牢性、信頼性と安全性**：VectorDB内のノイズの多いデータや不完全なデータを処理できる堅牢な検索アルゴリズムを実装し、一貫性のある正確な検索結果を提供します。
4. **透明性と説明可能性**：特定の情報が検索された理由と、それがクエリとどのように関連しているかを説明することで、検索プロセスの透明性を高めます。
5. **説明責任と是正可能性**：検索プロセスの性能を監視し、検索イベントをログに記録して、検索プロセスの監査と分析を可能にします。
6. **プライバシーとセキュリティ**：アクセス制御や暗号化機構を実装することで、VectorDBの機密性と完全性を保護します。検索プロセスを実行している間に、誤って機密情報を漏洩しないようにします。
7. **遵守性と制御可能性**：著作権制限を尊重し、不適切または違法なコンテンツの検索を回避するなど、検索プロセスが規制とポリシーを確実に遵守します。
8. **倫理性と公正性（バイアスの無い）**：VectorDBを慎重に管理し、差別を避ける公平性を考慮した検索技術を採用することで、検索プロセスにおける潜在的なバイアスを軽減します。
9. **環境的持続可能性**：エネルギー効率を高めるために検索アルゴリズムとデータ保存機構を最適化し、VectorDBの保存とアクセスによる環境への影響を最小限に抑えます。

データ保存

1. **人中心と人との整合性**：VectorDBは、倫理的なデータ ガバナンス ポリシーを遵守しながら、多様な視点を反映した、ユーザーのニーズと価値観に関連のある高品質のデータを保存する必要があります。
2. **アクセス性と包摂性**：VectorDB内のデータは、幅広い互換性のある形式で、承認されたユーザーが簡単に見つけて、アクセスできるようにし、包括性と多様な知識表現を促進する必要があります。
3. **堅牢性、信頼性と安全性**：データの完全性、正確性、および潜在的な脅威や中断に対する回復力を確保するために、データ検証、品質管理、バックアップおよびリカバリのメカニズム、および堅牢なセキュリティ対策を実装します。
4. **透明性と説明可能性**：明確なデータの出所と系統を維持し、包括的なメタデータとドキュメントを提供し、データの監査とログ記録を実装して、保存されたデータの透明性と理解を促進します。
5. **説明責任と是正可能性**：データ監視およびアラート システムを確立し、データの使用パターンを追跡し、データ管理の明確な責任範囲を定義して、説明責任を確保し、問題の特定を容易にします。
6. **プライバシーとセキュリティ**：データ暗号化、きめ細かなアクセス制御機構、およびデータの匿名化を採用して、機密データを不正アクセスから保護し、プライバシーを守ります。

7. **遵守性と制御可能性**：データ保存方法が関連規制（GDPR、HIPAAなど）に準拠していることを確認し、適切なデータ保持および廃棄ポリシーを実装し、包括的なデータ ガバナンス フレームワークを確立します。
8. **倫理性と公正性（バイアスの無い）**：データ パイプラインにバイアス検出および軽減機構を実装し、公正なデータ原則を遵守し、バイアスを最小限に抑えて公平性を促進するための倫理的なデータ収集と使用方法の実践を確保します。
9. **環境的持続可能性**：効率性を高めるためにデータ ストレージを最適化し、持続可能なデータ センターの実践を活用し、データ ライフサイクル管理戦略を実装して、ストレージ スペース、エネルギー消費、およびデータ保存の環境への影響を最小限に抑えます。

生成AIモデル

この項目は、上記シンプルなAIアプリケーションと変わりません。

RGAFシステム スコア

システム全体を評価する手段が必要です。システムがRGAFの視点にどの程度準拠しているかを表すスコアを開発することができます。スコアリング/評価機構は本ドキュメントの範囲外です。しかし、単純なアプローチとしては、AIシステムアーキテクチャの各コンポーネントを各視点によりスコアを付けて評価し、この評価値を平均化して全体のスコアを表すことが考えられます（加重平均はビジネス上の問題、時間および空間の複雑さによって異なります）。より詳細なアプローチとしては、ユーザー調査、専門家の評価、システム ログ、エネルギー消費量のモニタリング、その他のソースからデータを収集し、システム全体を表す加重スコアを割り当てることが挙げられます。

図5：RGAFのスコアリング機構の例

7. 結論と今後の活動

生成AIと大規模言語モデル（LLM）を原動力とする第五次産業革命は、効率性、洞察、そして機会において比類のない進歩をもたらし、私たちの生活に革命をもたらすでしょう。しかしながら、これらの技術は、ディープフェイクや誤情報の拡散、そして責任ある使用に伴う複雑さなど、重大な課題ももたらします。この革命は、IoT、人工知能、自動運転車、個別化医療といった変革を通じて、コンピューティングとインターネットベースの技術を日常生活に統合した第四次産業革命の基盤の上に築かれています。

生成AIの登場は、テキスト、画像、動画など、多様なコンテンツ形式を前例のない規模で作成できるようにすることで、これらの変革をさらに推し進めています。この機能はAIの潜在的な用途を拡大する一方で、公平性や堅牢性といった既存の課題を深刻化し、有害なコンテンツの生成、知的財産権に関する懸念、プライバシーの問題、信用性といった新たな複雑さをもたらします。これらの複雑さは、特にLLMや拡散モデルといった強力な生成モデルの急速な成長と広範な採用を考えると、責任あるAI（RAI）のためのフレームワークを大幅に拡張する必要性を浮き彫りにしています。

これらの基礎モデルは、膨大なWebクロールされたデータセットでトレーニングされるため、責任あるAIに対するユーザーの視点は多様化しています。

有害性や安全性などの特性は、文脈に応じたガバナンスの重要性をさらに強調しています[101]。責任あるAI開発への統一された堅牢なアプローチは、これらの技術を変革的なものにする創造性、流暢さ、一般性を維持しながら、これらの課題に対処するために不可欠です。

本研究では、オープンソース システム向けの責任ある生成AIフレームワーク（RGAF）の9つの視点を特定し、それらを世界の主要なイニシアチブ（米国NIST、EU AI ACT、シンガポールAI戦略、中国AI戦略など）と関連付けました。これらの9つの視点は、ニーズ、空間、時間に基づいて非常にバランスが取れています。

- 人中心と人との整合性
- アクセス性と包摂性
- 堅牢性、信頼性と安全性
- 透明性と説明可能性
- 説明責任と是正可能性
- プライバシーとセキュリティ
- 遵守性と制御可能性
- 倫理性と公正性（バイアスの無い）
- 環境的持続可能性

また、私たちは、LF-AI and DataのGenerative AI CommonsのModel Openness Frameworkとこれらの視点を関連付けました。典型的なAIアーキテクチャ間の関係性、そしてこれらの視点がアーキテクチャのどの部分に適用、実装できるかを探りました。この研究は、長い道のりの出発点です。今後、以下の取り組みを継続していく予定です。

- 責任あるAIフレームワークのビジネスへの影響
- AIの安全性とAIのセキュリティを連携させて責任あるAIを実現する方法
- RGAFの各視点をサポートするオープンソース プロジェクトの特定

RGAFの視点と評価指標およびツールにマッピングし（付録C：責任あるAI評価ツールと手法のサンプル リスト）、結論として、責任を生成AIソリューションに組み込むことは、必要不可欠であるだけでなく、これらの画期的なスマート テクノロジーが社会と個人の幸福を向上させるための基盤となります。RGAFの導入を通じて透明性、説明責任、そして倫理的開発を促進することで、私たちはこの新しい時代の複雑さを乗り越え、すべての人々の利益のために、生成AIの潜在能力を最大限に引き出すことができるようになります。

今後、このホームページで紹介されているResponsible AI Pathways（責任あるAIの方向性）の開発を継続していきます。

8. 付録

8.1. 付録A：AIの定義

文献には実に様々なAI用語が溢れており、それらが責任あるAIとどのように関連しているかが論じられています。例えば、倫理的AI、グリーンAI、安全なAI、責任あるAIなどです。以下に示すように、これらの用語は相互に関連しており、しばしば重複しますが、責任ある、倫理的かつ持続可能な方法でAIを開発・展開するというより広い文脈の中で、それぞれに固有の重点と適用範囲があります。

- **倫理的AI**：倫理的AIとは、道徳的原則と価値観に沿った方法で人工知能システムを開発・利用することを指します。AIアプリケーションが公平性、透明性、説明責任、そして人権尊重を考慮した設計・展開の保証がなされていることを意味します。
- **グリーンAI**：グリーンAIは、人工知能の環境への影響に焦点を当てています。エネルギー効率が高く、環境面で持続可能なAIシステムとアルゴリズムの開発に取り組んでいます。AIシステムに関連する二酸化炭素排出量と資源消費を最小限に抑えることを目指しています。
- **安全なAI**：安全なAIとは、危害を防止することに主眼に置いた人工知能システムの設計と実装を指します。これには、敵対的な攻撃に対する堅牢性、システムの信頼性の確保、意図しない結果を回避するための安全メカニズムの実装が含まれます。
- **持続可能なAI**：持続可能なAIとは、長期的な環境、経済、社会の持続可能性を考慮した人工知能の開発と展開を指します。これは、（グリーンAIにおける）エネルギー効率にとどまらず、AIが持続可能な開発に及ぼす全体的な影響と貢献について、より広い視点を包含するものです。
- **信頼できるAI**：信頼できるAIとは、人工知能システムへの確信性と信頼性を確立することを指します。これには、透明性、説明責任、そして信頼性と正確性を兼ね備えた結果の一貫した提供を確保することにより、ユーザー、利害関係者、そして一般の人々の間で信頼を築くことが含まれます。
- **信用できるAI**：信用できるAIは信頼できるAIに似ており、人工知能システムの信頼性と完全性を重視します。これは、AIアプリケーションが期待通りに、偏見なく、倫理的および法的基準に準拠した方法で動作することを保証することを含みます。
- **倫理的AI vs. 責任あるAI**：倫理的AIは特に道徳原則に焦点を当てていますが、責任あるAIは法的、社会的、文化的視点を含む、より広範な考慮事項を網羅しています。
- **グリーンAI vs. 持続可能なAI**：グリーンAIは、主にエネルギー効率などの環境への影響に重点を置いていますが、一方、持続可能なAIは、環境、経済、社会全体への長期的な影響を考慮しています。
- **安全なAI vs. 信用できるAI**：安全なAIは主に危害を防ぎ、システムの信頼性を確保することに重点を置いていますが、信用できるAIは透明性と説明責任を通じて確信性と信頼性を構築することに重点を置きます。
- **責任あるAI**：責任あるAIとは、AIを、責任を持ち、説明責任を果たしながら開発・利用するという、より広範な概念を指します。これには、倫理的な配慮だけでなく、法律、社会、文化的な視点への対応も含まれます。また、責任あるAIは、AIが社会に及ぼす潜在的な影響を認識する必要性を、企業や開発者に強調するものです。

そうです、それは妥当な概念化です。責任あるAIとは、倫理的AI、グリーンAI、AIの安全性、持続可能なAI、信頼できるAIなど、様々な視点と考慮事項を包含する包括的なハイレベル用語ととらえることができます。責任あるAIは、倫理、環境、安全性、持続可能性、そして信頼に関する視点を考慮し、人工知能の開発と展開において包括的かつ総合的なアプローチをとる必要性を強調しています。これは、AIシステムがより広範な社会的価値観や目標に沿った方法で開発・利用されることを保証するというコミットメントを反映しています。

8.2. 付録B：AI利害関係者のサンプル リスト

AI利害関係者とは、AIシステムの開発、展開、または利用に関心、影響力、または影響を与える個人、グループ、または組織を指します。欧州委員会のAI高等専門家グループ（EU Commission's High-Level Expert Group）は、AI利害関係者を「市民、消費者、企業、公的機関、研究者、市民社会組織、その他の関連主体を含むAIの開発、展開、および利用に関心を持つすべての人々」と定義しています[102]。世界経済フォーラム（World Economic Forum）は、AI利害関係者を「人工知能の開発、展開、および利用の影響を受ける、または関心を持つ個人、グループ、または組織」と定義しています[103]。スタンフォード哲学百科事典（Stanford Encyclopedia of Philosophy）は、AI利害関係者を「AIシステムの開発、展開、および利用に既得権益を持つ人々」と定義しています[104]。

このことは、以下の幅広いアクターを含みます。

- **システム製作者**：AIシステムまたはコンポーネントを開発します。例えば、学術界または産業界の機械学習研究者やエンジニアなどが挙げられます。
- **法律専門家**：AIシステムの法的視点を処理し、規制および知的財産法の遵守を確保します。
- **規制当局**：AIシステムを規制する規則を作成または編集する機関で、通常は、政府の政策を立案します。特に米国、EU、シンガポールでは、規制当局の関与が進んでいます。
- **システム ユーザー**：AIシステムを使用、改変、または研究します。例えば、AIエンジニア、健康研究者、教育研究者などが挙げられます。
- **エンドユーザー**：AIシステムの出力を、システムを変更したり、調査したりすることなく消費します。例えば、チャットボットを使用する個人や、コンテンツを生成するアーティストなどが挙げられます。
- **影響を受ける人**：例えば、自動化された意思決定など、AIシステムの出力によって直接的な相互作用無しに影響を受けます。擁護団体には、ACLU（アメリカ自由人権協会）やAlgorithmic Justice Leagueなどの擁護団体が含まれます。
- **データ提供者**：AIシステムの学習、テスト、または検証に使用されるデータを提供します。これには、特定の契約に基づいてデータを提供する機関、企業、または個人が含まれます。
- **インフラストラクチャ提供者**：AIシステムの開発と展開に必要なハードウェア、クラウド サービス、または計算リソースを提供します。
- **監査人/遵守責任者**：AIシステムが倫理ガイドライン、法的基準、性能測定基準に準拠しているかどうかを独立してレビューします。
- **倫理学者**：AIの開発と実装が倫理基準に準拠していることを確認し、AIの社会的影響を調査します。
- **セキュリティ専門家**：敵対的攻撃やデータ侵害などの悪意のある攻撃からAIシステムを安全に保つことに重点を置きます。
- **投資家/資金提供者**：多くの場合、株式または利益と引き換えに、AIシステムの開発に資金を提供します。

- **擁護団体/非営利団体**：デジタル権利、AI倫理、環境への影響など、より広範な社会的利益を代表する組織です。
- **教育者およびトレーナー**：AIシステムの作成、使用、または操作方法を人々に教えることに重点を置く個人または組織です。

8.3. 付録C：責任あるAI評価ツールと手法のサンプル リスト

視点	名称	簡単な説明
倫理性と公正性（バイアス）	TensorFlow's Responsible AI Toolkit	バイアスを特定、軽減し、プライバシーを保護し、透明性を促進します。
倫理性と公正性（バイアス）	IBMのAI Fairness 360	AIモデルの公平性を測定し、バイアスを軽減します。
倫理性と公正性（バイアス）	等価オッズ（Equalized odds）	モデルの予測が様々な属性に基づくグループ間で同様に正確であることを確認することで公平性を測定します。
倫理性と公正性（バイアス）	人口統計的均等性（Demographic parity）	モデルの予測が様々な人口統計グループ間で同様に分布しているかどうかを確認します。
倫理性と公正性（バイアス）	統計的同等性差異（Statistical Parity Difference）	多数派または特権を持つグループと少数派または不利な立場にあるグループの間で好ましい結果が生じる可能性の差を測定する公平性の指標です。
倫理性と公正性（バイアス）	差別的影響（Disparate Impact）	AIアルゴリズムが表面的には中立に見える場合でも、特定の人種、性別、民族などの保護された属性に基づいて、一部のグループに不釣り合いな不利益をもたらす可能性があります。この影響は、アルゴリズムを訓練する際に用いられたデータの偏りに起因し、意図しない差別につながることがあります[105]。
透明性と説明可能性	ローカル解釈可能なモデル不可知論的説明（LIME：Local Interpretable Model Explanations）	決定に寄与する最も重要な特徴を強調して、個々の予測を説明します。
透明性と説明可能性	シャプレー加法説明（SHAP：SHapley Additive exPlanations）	各機能に貢献スコアを割り当てることで、モデルの動作のグローバル ビューを提供します。

視点	名称	簡単な説明
倫理性と公正性（バイアス）	Microsoft Responsible AI Toolbox	モデルの公平性を評価し、予測に関する洞察を提供し、情報に基づいた意思決定を可能にします。
倫理性と公正性（バイアス）	IBM AI Explainability 360	モデルがどのように予測を行い、バイアスを特定するかを説明します。
倫理性と公正性（バイアス）	Amazon SageMaker Clarify	バイアスを検出し、より公平な結果を得るためにモデルの決定を説明します。
倫理性と公正性（バイアス）	Google's What-If Tool	モデルの行動を分析することで透明性を高め、公平性を向上させます。
倫理性と公正性（バイアス）	Fairness Indicators by TensorFlow	ユーザー グループ間のモデル性能を評価し、差異を特定します。
倫理性と公正性（バイアス）	IBMのAI Fairness 360	AIモデルの公平性を測定し、バイアスを軽減します。
倫理性と公正性（バイアス）	Ethics & Algorithms Toolkit by PwC	AIリスクを管理し、ガバナンス、遵守、リスク管理の全体にわたって倫理基準を確保します。
倫理性と公正性（バイアス）	DrivenDataのDeon	データ サイエンス プロジェクトに倫理チェックリストを追加し、説明責任と透明性を促進します。
倫理性と公正性（バイアス）	Ethical OS Toolkit	潜在的なリスクと社会的損害を特定し、倫理的行動の戦略を策定します。
性能指標	GLUE	自然言語処理（NLP）モデルの性能を評価するための一般言語理解評価ベンチマークです。
性能指標	HumanEval	プログラミング タスクを解決するコードを生成するAIシステムの性能を評価するためのベンチマークです。
性能指標	RealToxicityPrompts	有害な反応の生成に対するLMの堅牢性を評価するための10万個の共通プロンプトの集合です。

視点	名称	簡単な説明
性能指標	MLPerf	AIの幅広い分野におけるAI性能を評価するための、業界をリードするAIベンチマークです。学術界、研究機関、産業界のAIリーダーで構成されるコンソーシアムで、MLCommonsが主催しています。ベンチマーク対象となるタスクには、画像分類、物体認識、医用画像分析、自然言語処理（NLP）、レコメンダー システム、大規模言語モデル、ステーブル ディフュージョンなどがあります。
性能指標	TPCx-AI	あらゆる業界をリードするベンチマーク企業の一つであるTPCが開発した標準規格に基づく、エンドツーエンドのAIベンチマークです。他のベンチマークは機械学習トレーニングにおける計算負荷の高い部分にのみ焦点を当てる傾向がありますが、TPCx-AIはより包括的なアプローチを採用し、データの読み込み、データの前処理とラベル付け、スコアリングとサービングといったステップを網羅しています。
性能指標	MuJoCo (Multi-Joint dynamics with Contact)	ロボット、バイオメカニクス、グラフィックス、アニメーション、機械学習などの物理システムの制御における強化学習（RL：Reinforcement Learning）アルゴリズムの性能を評価するためのベンチマークです。2009年にワシントン大学によって開発されました。MuJoCoで実行されるプロジェクトでは、通常、リアルタイムよりも高速なシミュレーションにおいて、厳格な精度と安定性の要件を課した物理シミュレーションが用いられます。
性能指標	DAWNBench	さまざまな最適化戦略、モデル アーキテクチャ、ソフトウェア フレームワーク、ハードウェアにわたる精度、レイテンシ、コストを考慮して、トレーニングおよび推論タスクにおけるディープ ラーニング モデルの性能を評価するためのベンチマークです。
透明性と説明可能性	Trust LLM, Decoding Trust	AI評価、ベンチマーク、説明可能性ツールの例です - Mlcommons AI safety benchmark、Helm Benchmark Stanford、LM Evaluation Harness、Trulens：LLM 評価ツール、SHAP：説明可能なAIツール、QII：定量的入力影響。
データ品質チェック	データ プロファイリング ツール (Data profiling tools)	データの分布を分析し、外れ値を特定し、欠損値をチェックして、データの品質と代表性を確保します。

視点	名称	簡単な説明
データ品質チェック	データ出自追跡 （Data provenance tracking）	データの系統を監視して、モデル内でデータがどのように収集、変換、使用されるかを理解します。
モデルの監視と ドリフト検出	コンセプト ドリフト検出アルゴリズム	モデルの性能を時間の経過とともに監視し、データ分布の変化に応じてモデルの再トレーニングによる変更が必要になるタイミングを特定します。
人中心と 人との整合性	エキスパート レビュー （Expert review）	専門家がモデル出力をレビューして、精度を評価し、潜在的な問題を特定します。
人中心と 人との整合性	ユーザー フィードバックの仕組み	ユーザーからのフィードバックを収集して、潜在的なバイアスや改善領域を特定します。
性能指標	正確性	正しい予測の割合を測定する指標で、AIの性能の向上を示します。
性能指標	適合率 （precision）	肯定的な予測の精度を定量化します。精度は、誤検知のコストが高い場合に有効な指標です。例えば、メールのスパム検出において、誤検知とは、スパムではないメールをスパムと判定することです[106]。
性能指標	F値 （F1-score）	適合率（precision）と再現率（recall）の調和平均であり、両方の指標の重要性のバランスを考慮したものです。F1スコアは0～100%の範囲の値で、スコアが高いほど分類器の品質が高いことを示します。完璧なモデルのF1スコアは1になります[107]。
性能指標	再現率 （Recall）	再現率は、モデルが正しく識別した実際の肯定的ケースの割合です。感度（sensitivity）または真陽性率（positive rate）とも呼ばれています。再現率は、真陽性数を陽性例の総数で割ることで計算されます。陽性例の総数には、真陽性と偽陰性の両方が含まれます[108]。
性能指標	ROC-AUC （Receiver Operating Characteristic Area Under the Curve）	ROC曲線（Receiver Operating Characteristic Curve）における曲線下の面積（AUC：Area Under the Curve）は、機械学習において、バイナリ分類モデルの性能を評価するために使用される統計指標です[109]。

8.4. 付録D：責任あるAIフレームワークとガイドライン[110]を策定した国と組織のサンプルリスト

1. Australia AI https://www.uts.edu.au/human-technology-institute/news/report launch-state-ai-governance-australia
2. Brazil AI policy https://accesspartnership.com/access-alert- brazils-new-ai-bill-a comprehensive-framework-for-ethical- and-responsible-use-of-ai-systems/
3. Canada Data and AI Act https://ised-isde.canada.ca/site/innovation-better canada/en/artificial-intelligence-and-data-act; https://ised isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act aida- companion-document
4. India AI http://www.meity.gov.in/writereaddata/files/DIA_Presentation%2009.03.2023%20Fin al.pdf
5. Israel https://www.gov.il/en/pages/ai_2023
6. Japan AI policy https://www.meti.go.jp/shingikai/mono_info_ service/ai_shakai_jisso/pdf/20220128_2.pdf
7. New Zealand https://www.privacy.org.nz/publications/guidance-resources/ai/
8. Saudi Arabia https://www.whitecase.com/insight-our-thinking/ai-watch-global regulatory-tracker-saudi-arabia
9. Singapore https://www.whitecase.com/insight-our-thinking/ ai-watch-global regulatory-tracker-singapore
10. South Korea https://www.whitecase.com/insight-our-thinking/ai-watch-global regulatory-tracker-south-korea
11. Switzerland AI policy https://www.dsi.uzh.ch/en/esearch/projects/strategy lab/strategy-lab-21.html
12. The world economic forum https://intelligence.weforum.org/topics/a1Gb0000000pTDREA2/key issues/a1Gb00000017L8jEAE
13. United Arab Emirates https://www.whitecase.com/insight-our-thinking/ai-watch global-regulatory-tracker-uae
14. United Kingdom AI policy https://legalnodes.com/article/uk-ai-regulations
15. United States https://www.whitecase.com/insight-our-thinking/ai-watch-global regulatory-tracker-united-states

8.5. 付録E：責任あるAIフレームワークとガイドラインを策定した民間企業のサンプルリスト

1. Amazon https://aws.amazon.com/ai/responsible-ai/
2. Chat GPT https://chatgpt.com/
3. Deloitte https://www2.deloitte.com/us/en/pages/consulting/ articles/responsible use-of-generative-ai.html
4. FaceBook https://ai.meta.com/blog/facebooks-five-pillars- of-responsible-ai/
5. Google https://ai.google/responsibility/responsible-ai- practices/
6. HPE https://www.hpe.com/in/en/solutions/artificial- intelligence/ethics.html
7. IBM https://www.ibm.com/docs/en/watsonx-as-a-service?topic=ai-risk-atlas
8. IBM https://www.ibm.com/impact/ai-ethics
9. Info Tech Research Group https://www.infotech.com/sem/lp4/govern-the-use-of-ai responsibly-with-a-fit-for-purpose- structure
10. McKinsey https://www.mckinsey.com/capabilities/quantumblack/how-we-help clients/generative-ai/responsible-ai-principles
11. Microsoft https://www.microsoft.com/en-us/ai/principles- and-approach
12. PWC https://www.pwc.com/gx/en/issues/data-and-analytics/artificial intelligence/what-is-responsible-ai/responsible-ai- practical-guide.pdf
13. Snowflake https://www.snowflake.com/data-ai-predictions

8.6. 付録F：関連する責任あるAI活動と組織

AIの責任ある開発と利用を可能にするため、国際標準化機構（ISO）は最近、AI管理システムの新しい規格であるISO/IEC 42001を発表しました。ISOによれば、この規格は「技術が急速に進化する中でも、組織がAIを、責任を持って効果的に活用するために必要な包括的なガイダンスを提供する」とのことです。

1. ISO/IEC 20546:2021 - Big data - Overview and vocabulary
2. ISO/IEC 30106:2017 - Big data reference architecture
3. ISO/IEC TR 38505:2017 - Governance of data - Principles and practices
4. ISO/IEC TR 5072:2014 - Information technology - Vocabulary - Part 5: AI
5. ISO/IEC 2382-28:2017 - Vocabulary - Part 28: Artificial intelligence
6. ISO/IEC 24748-1:2020 - ISO/IEC 24748-8:2020

   Lifecycle - Metrics Par 1-8
7. ISO/IEC AWI 42001 - Artificial intelligence - Management system
8. ISO/IEC AWI 42002 - Artificial intelligence - Use cases
9. ISO/IEC AWI 42003 - Artificial intelligence - Trustworthiness
10. ISO/IEC AWI 42004 - Artificial intelligence - Risk management
11. ISO/IEC AWI 42005 - Artificial intelligence - Ethics and societal considerations
12. ISO/IEC 5962:2021 - ISPDX Specification (SBOM)
13. ISO/IEC 42001 - Artificial intelligence -Management System

IEEEで検討すべきもの

1. IEEE P7001 - Transparency of Autonomous Systems:
2. IEEE P7002 - Data Privacy Process:
3. IEEE P7003 - Algorithmic Bias Considerations:
4. IEEE P7006 - Personal Data Artificial Intelligence (AI) Agent:
5. IEEE P7007 - Ethically Driven Robotics and Automation
6. IEEE P7009 - Fail-Safe semi-autonomous systems.
7. IEEE P7010 - Wellbeing Metrics Standard for Ethical AI Autonomous Systems
8. IEEE P7011 - Process of Identifying and Rating the Trustworthiness in AI
9. IEEE P7014 - Ethical Considerations in AI and Autonomous Systems
10. IEEE P2863 - AI Governance

Generative AI CommonsのResponsible AI Workstreamは、NIST AI Safety Instituteコンソーシアムと連携しています。NIST AI Safety Instituteコンソーシアム（AISIC）は2024年2月8日に設立が発表されました。このコンソーシアムは200以上の組織を結集し、AIの測定とポリシーに関する科学的かつ実証的なガイドラインと標準を策定し、世界中のAIの安全性の基盤を築いています。コンソーシアムは以下のワーキング グループを運営しています。

1. Working Group #1：Risk Management for Generative AI
2. Working Group #2：Synthetic Content
3. Working Group #3：Capability Evaluations
4. Working Group #4：Red-Teaming
5. Working Group #5：Safety & Security

9. 謝辞

本研究にご参加いただき、責任ある生成AIソリューションの開発と展開の礎となるこのフレームワークの開発やフィードバックをしてくださった皆様に感謝申し上げます。視点の草案を提出してくださったMatt White氏にも感謝申し上げます。また、ワシントン大学タコマ校ビジネスアナリティクス センターおよびミルガード経営大学院、そして国際サービス イノベーション専門家協会（ISSIP：International Society of Service Innovation Professionals）からもご支援、ご激励、そして貴重なフィードバックを賜りました。本研究は、Demirkan教授のAmazonにおける役職とは関係ありません。

この記事は次のように参照してください。

Demirkan, Haluk; Zaalouk, Adel; Bhattacharya, Suparna and Malaika, Susan (2025) “The LF AI & Data Generative AI Commons Responsible Generative AI Framework (RGAF),” Linux Foundation AI and Data Group Generative AI Commons, 1-47, March 2025.

10. 脚注

1 https://www.psychologytoday.com/us/blog/the-digital-self/202310/the-5th-industrial-revolution-the-dawn-of-the-cognitive-age

2 Demirkan, H. and Spohrer, S. (2018) "Cultivating T-Shaped Professionals in the Era of Open Collaborative Innovation & Digital Transformation," *Informs Service Science Journal,* 10 (1), 98-109.

3 https://www.weforum.org/agenda/2023/12/ai-regulation-open-source/

4 Kalam Siddike, A., Spohrer, J., Demirkan, H. and Kohda, Y. (2018) "A Framework of Enhanced Performance: People's Interactions with Cognitive Assistants," Special Issue on Evolving and Innovating Big Data Systems Components: Expanding Analytics, Techniques, Services and Impacts, International Journal of Systems and Service-Oriented Engineering (IJSSOE), 8 (3), 1-17.

5 https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/navigating-the-generative-ai disruption- in-software

6 https://www.linuxfoundation.org/hubfs/LF%20Research/GenAI_Report_2023_011124

pdf?hsLang=en

7 https://openvoicenetwork.org/wp-content/uploads/2022/10/Ethical-Guidelines-for-Voice- Experiences-v2.0-2023.03.28.pdf

8 https://www.amazon.science/blog/responsible-ai-in-the-generative-era

9 https://blog.google/technology/safety-security/ an-update-on-our-child-safety-efforts-

and-commitments/

10 https://support.tiktok.com/en/using-tiktok/ creating-videos/ai-generated-content

11 https://newsroom.tiktok.com/en-us/partnering-with-our-industry-to-advance-ai-transparency-and-literacy

12 https://www.ftc.gov/policy/advocacy-research/ tech-at-ftc/2023/10/consumers-are-voicing- concerns-about-ai

13 https://medium.com/@oferher/harmonizing-creation-and-compensation-the-future-of-open-source-ai-in-a-copyrighted-world 81c2831fcd41)

14 https://www.nist.gov/system/files/documents/2024/02/15/ID012%20-%202024-02-01%2C%20Thorn%20and%20ATIH%2C%20 Comments%20on%20AI%20EO%20RFI.pdf

15 https://www.forbes.com/councils/forbestechcouncil/2023/11/20/the-pros-and-cons-of-using-synthetic-data-for-training-ai/

16 https://www.ftc.gov/news-events/news/press-releases/2022/06/ftc-report-warns-about-using-artificial-intelligence-combat-online problems

17 https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_ identities_0.pdf

18 Moghaddam, Y., Yurko, H., Demirkan, H., Tymann, N. and Rayes, A., The Future of Work: How Artificial Intelligence Can Augment Human Capabilities (in Series Collaborative Intelligence: People, AI, and the Future of Work,), Business Expert Press, March 2020.

19 https://www.ftc.gov/news-events/news/press-releases/2023/05/ftc-warns-about-misuses-biometric-information-harm-consumers

20 https://www.weforum.org/agenda/2023/03/why-businesses-should-commit-to-responsible-ai/

21 https://sites.google.com/view/responsible-gen-ai- tutorial

22 A Systematic Literature Review of Human-Centered, Ethical, and Responsible AI

23 A Perspective from Human- Computer Interaction

24 "Human-Centered Artificial Intelligence" by Ben Shneiderman

25 What is Human-Centered about Human-Centered AI? A Map of the Research Landscape

26 What is Human-Centered about Human-Centered AI? A Map of the Research Landscape

27 A comprehensive overview of barriers and strategies for AI implementation in healthcare: Mixed-method design | PLOS ONE

28 A-STUDY-OF-BARRIERS-AND-BENEFITS-OF-ARTIFICIAL-INTELLIGENCE-ADOPTION-IN-SMALL-AND-MEDIUM-ENTERPRISE.pdf

29 Overcoming Common Challenges in AI Adoption for Small Businesses

30 OECD research on AI for labour market accessibility: opportunities and challenges - AAATE

31 https://arxiv.org/pdf/2405.06624

32 https://arxiv.org/pdf/2405.06624

33 https://arxiv.org/pdf/2405.06624

34 https://mlcommons.org/benchmarks/ai-safety/general_purpose_ai_chat_benchmark/

35 https://www.semanticscholar.org/paper/LegoAI%3A-Towards-Building-Reliable-AI-Software-for-Hou Xu/07ada28d2c32c81eaab7aa7ee8ade7d43b63f334

36 Explainable Artificial Intelligence (XAI) (Archived)

37 IBM Artificial Intelligence Pillars - IBM Policy

38 https://arxiv.org/pdf/2302.05284

39 https://arxiv.org/pdf/2311.13158

40 https://arxiv.org/pdf/2311.13158

41 AI Technologies, Privacy, and Security - PMC

42 Intelligent Agents

43 Privacy and Security Implications of Cloud-Based AI Services : A Survey

44 OWASP Top 10 Security Vectors for LLM Applications: https://owasp.org/www-project-top-10-for-large-language-model applications/assets/PDF/OWASP-Top-10-for-LLMs-2023-v1_1.pdf

45 Format preserving encryption

46 What is Homomorphic Encryption? | IBM

47 Federated or Split? A Performance and Privacy Analysis of Hybrid Split and Federated Learning Architectures

48 [2312.12148] Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment

49 https://senhasegura.com/post/the-pillars-of-cybersecurity

50 https://kpmg.com/ch/en/insights/technology/artificial-intelligence-ensuring-compliance.html

51 https://consilium-europa.libguides.com/c.php?g=690732&p=4948483

52 https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1. pdf

53 https://aiverifyfoundation.sg/

54 https://www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence development-plan-2017/

55 https://link.springer.com/chapter/10.1007/978-3-031-40837-3_1

56 https://arxiv.org/pdf/2406.14758

57 https://arxiv.org/pdf/2406.14758

58 https://arxiv.org/pdf/2408.12599#page=1.00&gsr=0

59 Ethical approaches in designing autonomous and intelligent systems: a comprehensive survey towards responsible development | AI & SOCIETY

60 Moghaddam, Y., Demirkan, H. and Spohrer, J., T-Shaped Professionals: Adaptive Innovators (in Series: Service Systems and Innovations in Business Society), Business Expert Press, Aug 2018.

61 Ethical AI in practice: Balancing technological advancements with human values

62 https://ijsra.net/content/ethical-ai-practice-balancing-technological-advancements-human-values

63 https://theconversation.com/building-fairness-into-ai-is-crucial-and-hard-to-get-right-220271

64 [2305.03720] Training Is Everything: Artificial Intelligence, Copyright, and Fair Training

65 Good Models Borrow, Great Models Steal: Intellectual Property Rights and Generative AI

66 https://ijsra.net/content/ethical-ai-practice-balancing-technological-advancements-human-values

67 Is the future of AI sustainable? A case study of the European Union | Emerald Insight

68 https://arxiv.org/html/2406.05303v1

69 Beyond Efficiency: Scaling AI Sustainably

70 https://eudl.eu/doi/10.4108/eai.20-11-2021.2314097

71 https://arxiv.org/html/2406.05303v1

72 EU AI Act 2024/1689 - https://eur-lex.europa.eu/eli/reg/2024/1689

73 https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1. pdf

74 https://aiverifyfoundation.sg/

75 Full Translation: China's `New Generation Artificial Intelligence Development Plan' (2017)

76 https://www.nsfc.gov.cn/publish/portal0/tab442/info91118.htm

77 https://www.gov.cn/zhengce/zhengceku/202407/P020240702716282797987.pdf

78 https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm

79 http://www.cac.gov.cn/2019-04/23/c_1556061196841081.htm

80 https://www.chinalawtranslate.com/en/cybersecurity-law/

81 http://www.npc.gov.cn/englishnpc/Law/2021-08/27/content_2099222.htm

82 https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

83 https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker singapore#:~:text=The%20Monetary%20Authority%20of%20Singapore,firms%20to%20consider%20when%20using

84 https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1. pdf

85 https://artificialintelligenceact.eu/recital/27/

86 https://baai.ac.cn/en/ethical-norms-for-artificial- intelligence/

87 https://www.gov.cn/zhengce/zhengceku/202407/P020240702716282797987.pdf

88 [2403.13784] The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency, and Usability in Artificial Intelligence

89 https://www.youtube.com/watch?v=DEe1GttUWKI&list=PLbzoR-pLrL6oII-dLqw9vpifOU_kTycj6&index=29

90 Openwashing - Wikipedia

91 https://arxiv.org/pdf/2108.07258

92 https://arxiv.org/pdf/2306.11698

93 https://decodingtrust.github.io/

94 https://decodingtrust.github.io/leaderboard/

95 The Shift from Models to Compound AI Systems - The Berkeley Artificial Intelligence Research Blog

96 Function Calling - OpenAI API

97 The Shift from Models to Compound AI Systems - The Berkeley Artificial Intelligence Research Blog

98 https://www.ibm.com/think/topics/ai-agents

99 Beyond LLMs: Compounds Systems, Agents, and Whole AI Products

100 Zhou, L., Paul, S., Demirkan, H., Yuan, L., Spohrer, J., Zhou, M., & Basu, J. (2021). "Intelligence augmentation: Towards building human machine symbiotic relationship," AIS Transactions on Human-Computer Interaction, 13(2), pp. 243-264.

101 https://research.google/pubs/the-many-faces-of-responsible-ai/

102 https://ec.europa.eu/digital-single-market/en/ news/ethics-guidelines-trustworthy-ai

103 https://www.weforum.org/agenda/2019/05/artificial-intelligence-stakeholders-need-to-work-together/

104 https://plato.stanford.edu/entries/ethics-ai/

105 https://www.brookings.edu/articles/the-legal-doctrine-that-will-be-key-to-preventing-ai discrimination/#:~:text=As%20developments%20in%20artificial%20intelligence,impacts%20people's%20rights%20and%20opportunities

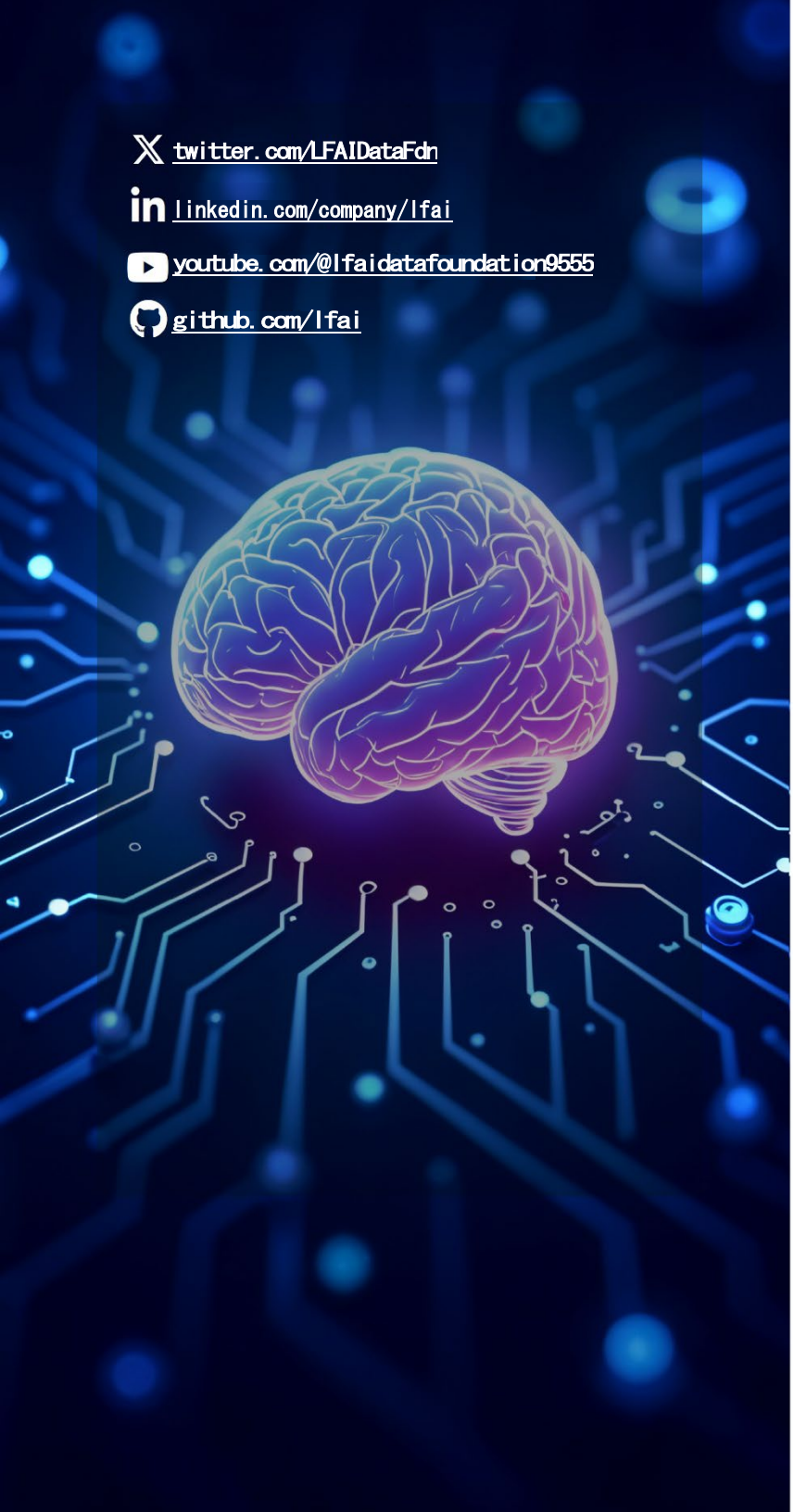
106 https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

107 https://arize.com/blog-course/f1-score/#:~:text=How%20Do%20Precision%20and%20Recall,of%20these%20metrics%20and%20 why

108 https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall#:~:text=balance%20precision:%20recall.-,What%20is%20recall?,common%20term%20in%20machine%20learning

109 https://www.pickl.ai/blog/auc-roc-curve- machine-learning/#:~:text=Machine%20Learning%20algorithms.-,Understanding%20AUC%20in%20Machine%20Learning,values%20indicating%20better%20model%20performance.

110 https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker



LF AI & DATA

LF AI & Data Foundationについて

LF AI & Dataは、Linux Foundation傘下の財団であり、人工知能（AI）とデータ分野におけるオープンソースのイノベーションを支援しています。LF AI & Dataは、オープンソースのAIとデータを支援し、オープンソース技術を用いたAIおよびデータ製品やサービスを容易に開発できる、持続可能なオープンソースAIとデータのエコシステム構築を目指して設立されました。LF AI & Dataは、オープンソース技術プロジェクトの調和と加速を支援するため、オープンガバナンスに基づく中立的な環境下でのコラボレーションを促進しています。LF AI & Data Foundationの現在のプロジェクトのポートフォリオをご覧ください。そして、LF AI & Data Foundationの下で、オープンソースAIまたはデータ プロジェクトのホスティングについてのご相談について、お気軽にお問い合わせください。

2025年3月

本訳文について

この日本語文書は、THE LF AI & DATA GENERATIVE AI COMMONS RESPONSIBLE GENERATIVE AI FRAMEWORK (RGAF) - V0.9の参考訳として、The Linux Foundation Japanが便宜上提供するものです。

翻訳協力：天満尚二